SAPIENZA UNIVERSITY OF ROME

DOCTORAL THESIS

Interpretable statistics for complex modelling: quantile and topological learning

Candidate:
Tullia PADELLINI

Supervisor:
Pierpaolo BRUTTI

Scuola di Dottorato "Scienze Statistiche"
Curriculum Methodological Statistics

Department of Statistical Sciences

February 2019

Table of Contents

Abstract

As the complexity of our data increased exponentially in the last decades, so has our need for interpretable features. This thesis revolves around two paradigms to approach this quest for insights.

In the first part we focus on parametric models, where the problem of interpretability can be seen as a "parametrization selection". We introduce a quantile-centric parametrization and we show the advantages of our proposal in the context of regression, where it allows to bridge the gap between classical generalized linear (mixed) models and increasingly popular quantile methods.

The second part of the thesis, concerned with topological learning, tackles the problem from a non-parametric perspective. As topology can be thought of as a way of characterizing data in terms of their connectivity structure, it allows to represent complex and possibly high dimensional through few features, such as the number of connected components, loops and voids. We illustrate how the emerging branch of statistics devoted to recovering topological structures in the data, Topological Data Analysis, can be exploited both for exploratory and inferential purposes with a special emphasis on kernels that preserve the topological information in the data.

Finally, we show with an application how these two approaches can borrow strength from one another in the identification and description of brain activity through fMRI data from the ABIDE project.

Acknowledgements

I am forever in debt with Håvard Rue for his help and his hospitality. I would also like to acknowledge Peter Congdon for providing the hospitalization for self harm data used in Section 1.7.1 and Bertrand Michel for referring to the Fullerenes analysed in Section 3.2.

Special thanks go to the reviewers Massimo Ventrucci and Larry Wasserman for their valuable comments and constructive feedback.

Introduction

This thesis is based on two main pillars: Quantile Regression and Topological Data Analysis.

Quantile Learning The first part of the thesis revolves around Quantile Regression, a supervised technique aimed at modeling the quantiles of the conditional distribution of some response variable. With respect to “standard” regression, which is concerned with modeling the conditional mean, Quantile Regression is especially useful when the tails of the distribution are of interest, as for example when the focus is on extreme behavior rather than average, or when it is important to assess whether or not covariates affect uniformly different levels of the population.

Even though the idea dates back to Galton (1883) (as noted in Gilchrist (2008)), Quantile Regression was formally introduced only relatively recently by Koenker and Bassett (1978). Since then, the use of quantiles in regression problems has seen an impressive growth and has been thoroughly explored in both the parametric (see Yue and Rue (2011), Wang, McKeague, and Qian (2017)), and non-parametric framework (see Yu and Jones (1998), Takeuchi et al. (2005), Li and Racine (2007)) with applications ranging from the Random Forest Quantile Regression of Meinshausen (2006) to D-vine copulas for quantiles in Kraus and Czado (2017), through Quantile Regression in graphical models as in Ali, Kolter, and Tibshirani (2016).

One of the most significant developments in the Quantile Regression literature has been the introduction of the Asymmetric Laplace Distribution (ALD) as a working likelihood Yu and Moyeed (2001). From a frequentist point of view, the use of the ALD gave rise to a class of likelihood based method for fitting quantile models and has been instrumental in introducing random effects in linear and non linear Quantile Regression models; see for example Geraci and Bottai (2007), Geraci and Bottai (2014), Geraci (2017) or Marino and Farcomeni (2015) for a more comprehensive review. The introduction of the ALD has been even more critical in the Bayesian framework, where the likelihood is required in inferential procedure Yu and Moyeed (2001). As a result, fully bayesian versions of Quantile Regression, such as the additive mixed Quantile model of Yue and Rue (2011), as well as Quantile Bayesian Lasso and Quantile Bayesian Elastic Net, have been developed in the last couple of years, examples being Alhamzawi, Yu, and Benoit (2012) or Li, Xiy, and Lin (2010). Extensions of the Asymmetric Laplace Distribution such as the Asymmetric Laplace Process (Lum and Gelfand (2012)), broadened Quantile Regression to spatially dependent data. Despite their popularity however, ALD based methods are not always satisfactory, especially in terms of uncertainty quantification. The use of the ALD introduces an unidentifiable parameter in the posterior variance, hence any inference beside point estimation is precluded (Yang, Wang, and He 2016).

We propose a model-based approach for Quantile Regression that considers quantiles of the generating distribution directly, and thus allows for a proper uncertainty quantification. We then create a link between Quantile Regression and generalized linear models by mapping the quantiles to the parameter of the response variable. This formulation not only recast Quantile Regression in a much more cohesive setting and overcomes the fragmentation that characterizes the Quantile Regression literature, but it is also key to an efficient and ready-to-use fitting procedure, as the connection allows to estimate the model using `R-INLA` (Rue, Martino, and Chopin (2009) and Rue et al. (2017)).

Additionally, we extend our model based approach in the case of discrete responses, where there is no 1-to-1 relationship between quantiles and distribution's parameter, by introducing continuous generalizations of the most common discrete variables (Poisson, Binomial and Negative Binomial) to be exploited in the fitting.

Topological Learning In the second part of the thesis we focus on Topological Data Analysis (`TDA`), a rapidly growing branch of statistics whose aim is estimating topological invariants of unobserved manifolds, typically (but not exclusively) through point–clouds sampled on them. `TDA` can be seen as a way of uncovering the "shape of the data" in terms of their topology. As topology is a rather broad definition of shape, which is focused on the connectivity structure, `TDA` methods are intrinsically related with clustering, making them a great exploratory tool for high dimensional and highly complex data.

`TDA` has a relatively short history, being based on Persistent Homology Groups, topological invariants introduced only at the beginning of 2000 by Edelsbrunner, Letscher, and Zomorodian (2002). The main tool of this class of methods, the Persistence Diagram, a topological summary containing both topological features and a measure of their importance, has been investigated from a statistical perspective even more recently (B. T. Fasy et al. 2014, Chazal, Glisse, et al. (2015)). While doing statistics on the Persistence Diagram has yield positive results (Chazal, Fasy, et al. 2015), statistics using the Persistence Diagram has proven to be more challenging, as even basic quantities such as the mean are not easy to compute or to interpret (Turner et al. 2014). For this reason several alternative representation of the Diagram, typically in the form of functional object, have been proposed (Bubenik (2015), Adams et al. (2017) or Moon, Giansiracusa, and Lazar (2018) to name a few examples).

We introduce a new topological summary for scale-spaces, the Persistence Flamelet, which allows to extend `TDA` to the case of object that have multiple resolution, such as time series, which depend on time, or smoothers in general, which involve a tuning parameter. We investigate in particular the case of kernel density estimators, where the scale parameter is the bandwidth, and we show how its topology changes with it. We prove that the Flamelet is not just a visualization tool by characterizing it probabilistically and showing that Central Limit Theorem and Law of Large Numbers hold for this new object.

Even though the impressive growth of `TDA` literature in the last couple of years has yield several inference–ready tools, this hype has not yet been matched by popularity in the

	Quantile Regression	Topological Data Analysis
Inferential Approach	Bayesian	Frequentist
Computation	Fast	Slow
Field	Traditional Statistics	Borderline Statistics

practice of data analysis, we thus focus on the potential of topological characterization as a Learning tool, with a special emphasis on Supervised problems. In order to perform inference using topological summaries, which are typically defined in spaces which are not amenable to direct modelling, we adopt a kernel approach to recast the learning into more familiar vector spaces. We define a *topological exponential kernel*, we characterize it, and we show that, despite not being positive semi-definite, it can be successfully used in regression and classification tasks. We examine in particular the former and we show how to use Persistence Diagram as covariate and as responses in regression problems. Finally, we show preliminary, yet encouraging, results of combining quantile methods with TDA to gain insights on brain activity. Building on functional connectivity, which in the literature has been analysed mostly with respect to its 0-dimensional structure, we show how TDA allows straightforwardly to investigate higher dimensional features as well, and then we use the quantile methods introduced in the first part of the thesis to better understand phenotypical determinants of the topological structure.

A tale of two thesis? At a first glance, there is little in common between the two topics themselves, as Quantile Regression is well established in "classical" statistics, while Topological Data Analysis is an emerging research area at the boundary between Statistics and Computational Topology. Our contributions also appear to go in opposite direction, since we approach Quantile Regression from a model-based perspective, heavily relying on parametric modelling to exploit fast and efficient Bayesian fitting procedure (INLA), while we opt for a parameter-free approach for Topological Inference, adopting classical tools in non-parametric statistics such as kernels. Finally, from a computational standpoint, thanks to INLA, we are providing extremely efficient and extremely fast implementations in the Quantile Regression setting, while, due to the indefiniteness of topological kernels, we turn to non-efficient solvers for classification problems using topological summaries, whose computation is already very time-consuming.

At a closer look however, the contribution presented in this thesis are all trying to pursue the same goal, that is *interpretable characterization of data.* Despite being at the core of learning, interpretability is in fact not a uniquely defined concept but there are many different declinations of this notion, depending on the task and most importantly on the information already available. Our contribution can be seen as as an attempt to enforce interpretability at the two end of the spectrum of model knowledge, i.e. the case where we know almost everything (that is, we have a parametric model for our data) and the case we know almost nothing (i.e. we don't even know the support of the data generating process).

From a parametric perspective, the interpretation typically goes through the parameters. Our model-based approach to Quantile Regression can be thought of a way of reparametrizing the model in terms of its quantiles. While the mean needs not to exist, the quantile are always defined and they retain the same interpretation regardless of the complexity of the model considered, hence a parametrization in terms of quantiles can be thought of as a "universal" parametrization.

The main argument against a parametrization based on quantiles may be that in the discrete case quantiles still exist but are not unique, which is why we focus on the discrete case, proposing a model-aware approach to overrule this objection.

As for the case where preliminary information is close to null, there may be even too many ways of finding a meaningful characterization of data, but topological invariants stand out for many reasons. The main one is of course their interpretability and the relevance of such interpretable objects in statistical learning: topological features of dimension 0, connected components, can in fact be thought of as clusters or peaks, while topological features of dimension 1, loops, represent periodic structures. Another advantage of a topological characterization is that it can be computed for most kind of data, from more standard point-clouds, to functional data or networks, which is especially appealing in the era of "complex data". Finally, a topological characterization does not depend on the coordinates of the data and it is rather robust with respect to deformation, which makes it very flexible.

Chapter 1

Model Based Quantile Regression

1.1 Motivation

Classical (mean) regression methods model the average behavior, which despite being an useful summary measure, does not capture how the covariates may not affect in the same way all levels of population. Quantile regression allows to quantify the effect of the covariates at each different quantile level, hence giving us a more complete picture of the phenomenon in analysis. As a motivating example to understand the use of this class of methods, let us consider data from the NBA $2016-2017$ season. For each of the 484 players in the NBA we consider the following variables:

- $\mathtt{Y}$: Points scored in the whole season
- $\mathtt{X}$: Minute played per game (on average)
- $\mathtt{E}$: Number of games played in the season

While in classical mean regression we would be interested in modeling conditional expectation $\mathbb{E}[Y|X=x]$, thus analyzing the behavior of the average player, a Quantile Regression model is concerned with the behavior of specific classes of players.

As opposed to (European) football, where only attacker scores, in basketball roles are not as well defined and all the players may score, hence usually, ceteris paribus, players that scores more are better players than those who score less. This implies that if α is the quantile level, then $\mathbb{Q}_\alpha(Y|X=x)$ models the level α player:

- $\alpha = 0.50$ median player
- $\alpha = 0.75$ good player
- $\alpha = 0.25$ bad player

We assume $\mathtt{Y|X} = \mathtt{x} \sim \text{Poisson}(\lambda)$ and that the level α quantile of the number of points scored depends on the minutes played, and adopt the following model for the level α quantile of the conditional distribution of the response variable $\mathbb{Q}_\alpha(\mathtt{Y}|\mathtt{X}=\mathtt{x})$:

$$\mathbb{Q}_\alpha(\mathtt{Y}|\mathtt{X}=\mathtt{x}) = \mathtt{E}\exp\{\beta_\alpha \mathtt{x}\}$$

where the exposure $\mathtt{E}$ is needed to take into account the fact that players that have played more games have more chances to score. As we can see from Table 1.1, the effect of the minutes played is very different among the different players' groups, and how it compares to the estimate for the average player.

Quantile level α	$\hat{\beta}_\alpha$	$\hat{\beta}_\alpha/\hat{\beta}_{\text{mean}}$
0.01	0.01	0.08
0.25	0.01	0.09
0.50	0.06	0.63
0.75	0.06	0.65
0.99	0.09	1.03

Table 1.1: Estimated β_α for different quantile levels α.

It is no surprise that the time played by an all star player is more valuable in terms of points scored, and, in fact, we an see that one minute played by a great player ($\alpha = 0.99$) is worth $\beta_{0.99}/\beta_{0.01} \approx 13$ times a minute played by a rather poor player ($\alpha = 0.01$). What may not be as obvious instead, is that the "average" player is not really representative, as the estimate of β for the mean model is closer to the franchise players than to that of the median player, which motivates us to explore regression methods beyond the mean.

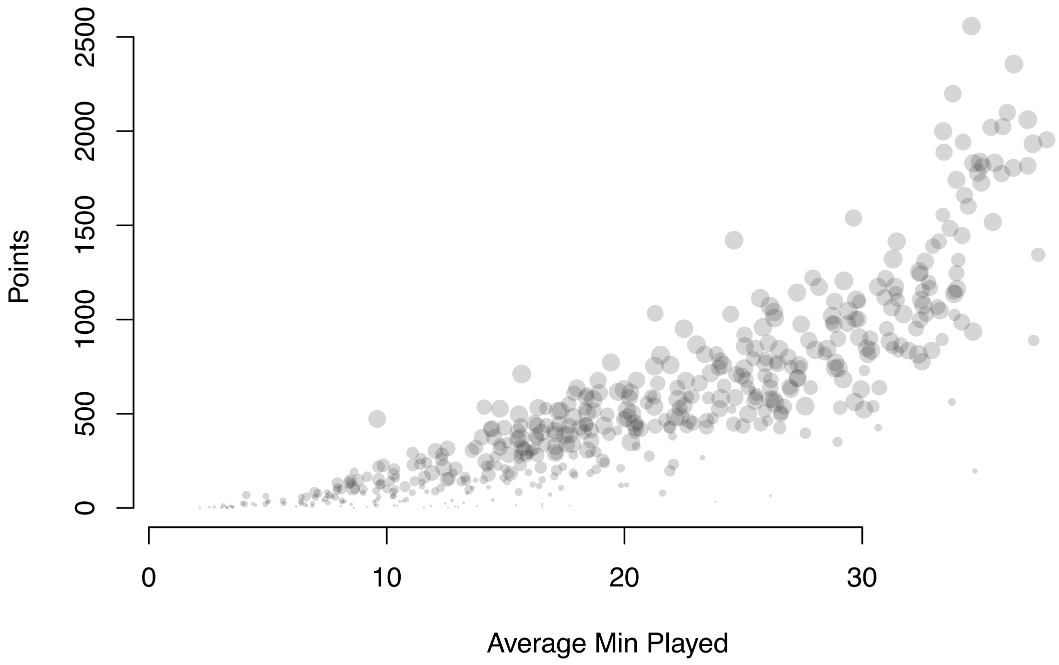

Figure 1.1: Minutes on the court vs Points scored per player, size proportional to the number of games played.

	Loss Function $L(Y, r(X))$	Regression Function $r^*(X)$
Quadratic Loss	$(Y - r(Y))^2$	$\mathbb{E}[Y\|X]$
$0 - 1$ Loss	$\mathbb{1}\{Y \neq r(X)\}$	$\text{Mode}(Y\|X)$
Absolute Loss	$\|Y - r(X)\|$	$\text{Median}(Y\|X)$
Check Loss	$(Y - r(X))(\alpha - \mathbb{1}\{Y - r(X) < 0\})$	$\mathbb{Q}_\alpha(Y\|X)$

Table 1.2: Most common loss functions and corresponding regression functions.

1.2 A Decision theory intermezzo

The broad goal of regression methods is to explain a random variable Y_i as a function of observed and/or latent covariates X; in formulas

$$Y = r(X) + \varepsilon \tag{1.1}$$

where ε is an error term which takes into account the randomness of the Y, while $r(\cdot)$, the *regression function*, represents the deterministic part of the relation between the response and the covariates. The regression function $r(X)$ is a summary of the conditional distribution of $Y|X$, chosen to minimize the expected loss (or risk) occurring when we neglect the error term to explain Y, or, in other words, the deterministic term $r(X)$ must be chosen so that, on average, it is "close" to the random variable Y. If the loss is taken to be the quadratic loss, i.e.

$$L(Y, r(X)) = (Y - r(X))^2$$

for example, then the regression function minimizing the risk is the conditional mean $\mathbb{E}[Y|X]$. A different loss function results in a different interpretation of the deterministic term of the regression, as shown in Table 1.2.

The choice of the check (or pinball) loss $\rho_\alpha(x) = x(\alpha - \mathbb{1}\{x < 0\})$, a tilted version of the absolute value, results in the regression function being the conditional quantiles.

$$\begin{aligned} r^*(X) &= \arg\min_{r(X)} \mathbb{E}[L(Y, r(X))] \\ &= \arg\min_{r(X)} \mathbb{E}[\rho_\alpha(Y - r(X))] \\ &= \mathbb{Q}_\alpha(Y|X). \end{aligned}$$

No loss function is uniformly better than the others, but each has different strengths. The advantage of the check loss over the quadratic loss (hence of Quantile Regression over mean regression), for example, is that it gives a more complete picture of the distribution of $Y|X$, it is more robust with respect to outliers, it allows for dealing with censored data without additional assumptions, and most importantly it allows to model extreme behavior.

pdf
2

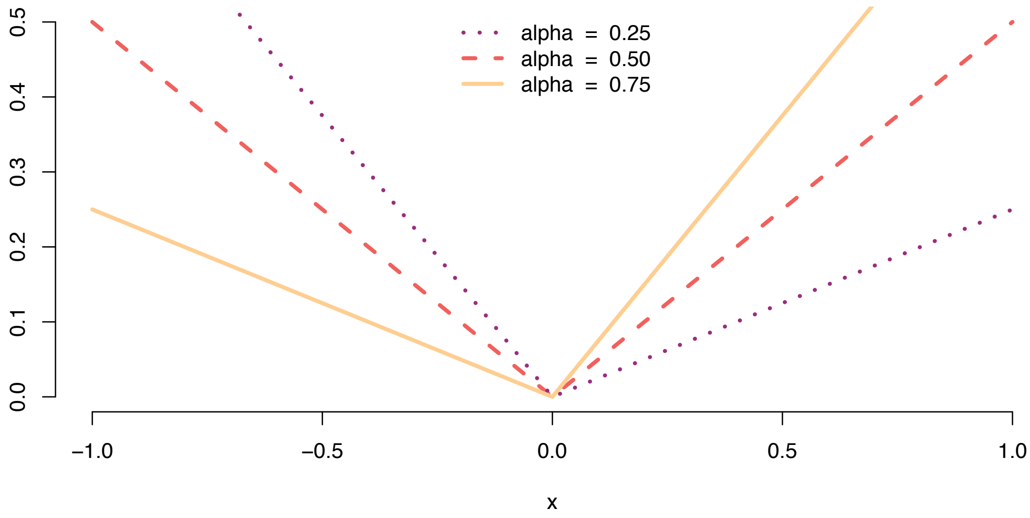

Figure 1.2: Check Loss

As in mean regression, Quantile Regression models can be parametric (which are the ones we will focus on), semi-parametric or non parametric altogether. In the first and most basic formulation of Koenker and Bassett (1978)}, quantile linear regression, the quantile of level α of the conditional distribution $Y|X$, can be modeled as:

$$\mathbb{Q}_\alpha(Y|X) = X^t\beta_\alpha$$

where the notation β_α highlights the dependence of the regression coefficients to the quantile level. Given a sample $\mathcal{D}_n = \{(Y_i, X_i)\}_{i=1}^n$, estimate for the regression coefficient β_α can be found by minimizing the empirical risk:

$$\begin{aligned} \widehat{\beta}_\alpha &= \arg\min_{\beta_\alpha} \widehat{\mathbb{E}}[\rho_\alpha(Y_i - X_i^t\beta_\alpha)] \\ &= \arg\min_{\beta_\alpha} \sum_{i=1}^n \rho_\alpha(Y_i - X_i^t\beta_\alpha). \end{aligned} \tag{1.2}$$

This is a standard linear programming (LP) problem and can be trivially solved by means of *simplex method* or *interior point methods* (Koenker and d'Orey 1987, Koenker and Ng (2005)).

1.3 Asymmetric Laplace Distribution

Quantile regression as defined by the optimum problem in Equation (1.2) does not require any distributional assumption for the response variable, and is thus an intrinsically non parametric (in the sense of model–free) method. This lack of generating model assumption implies that there is no likelihood, which is disturbing to some, especially Bayesians. In order to adopt likelihood based inferential procedure in the context of Quantile Regression, pseudo-likelihood approaches have been proposed. Among those, the one method dominating the literature is to exploit the Asymmetric Laplace Distribution (ALD) as a working likelihood, as first suggested by Yu and Moyeed (2001).

A random variable X has an Asymmetric Laplace Distribution with parameters α, scale σ and center μ, i.e. $X \sim \text{ALD}(\alpha, \sigma, \mu)$, if it has density

$$f_X(x) = \frac{\alpha(1-\alpha)}{\sigma} \exp\left\{-\frac{\rho_\alpha(x-\mu)}{\sigma}\right\}.$$

By assuming that the conditional distribution for the response variable $Y|X$ is an ALD with parameters α taken to be the quantile level we are interested in, and $\mu = X^t\beta$ (or a more complicated function of $X^t\beta$ if we want to move beyond linear Quantile Regression) the likelihood corresponding to a sample $\mathcal{D}_n = \{(Y_i, X_i)\}_{i=1}^n$ is:

$$L(\beta; \mathcal{D}_n) \propto \exp\left\{-\frac{\sum_{i=1}^n \rho_\alpha(Y_i - X_i^t\beta)}{\sigma}\right\}.$$

As the log-likelihood is proportional to minus the check function, it is immediate to see that the Maximum Likelihood Estimator (MLE) corresponds to the Quantile Regression estimator in Equation (1.2), hence the ALD translates Quantile Regression into a likelihood based estimation setting, without making true distributional assumption on the response variable; it is thus not the generating model but a working model (or likelihood).

Although most of what said in the following applies to the frequentist domain as well, as the presence of a generating model and hence a likelihood is especially critical for Bayesian procedures, from here on we will focus mostly on the Bayesian approach to Quantile regression.

1.3.1 Bayesian Pros of the ALD

From a theoretical point of view, one strong justification to the use of ALD in the Bayesian framework is the good behavior of estimates obtained adopting the ALD as a likelihood for data generated from a different distribution. This was shown empirically in Yu and Moyeed (2001) and then investigate more thoroughly by Sriram, Ramamoorthi, and Ghosh (2013), which proved the posterior consistency (as well as convergence rate to the true value) of ALD-based estimators for misspecified models under both proper and improper priors yielding proper posteriors. As a side result, this motivates flexibility in the choice of prior distribution in relation to the ALD. For example Yu and Moyeed (2001) proves that the posterior distribution is proper even with an improper uniform prior. Regularized version of Quantile Regression, such as LASSO, Elastic-net and SCAD have also been explored in the Bayesian framework exploiting the ALD as a working likelihood (Li, Xiy, and Lin 2010 Alhamzawi and Yu (2014)).

From a more practical point of view, the popularity of the ALD in the Bayesian framework stems mostly from the ease of implementation of the resulting fitting procedure. A random variable $Y \sim \text{ALD}(\alpha, \sigma, \mu)$ admits in fact the following decomposition:

$$Y = \sigma(\theta_1 V + \theta_2 Z\sqrt{V}) \tag{1.3}$$

where $\theta_1 = (1-2\sigma)/(\sigma - \sigma^2)$, $\theta_2^2 = 2/(\sigma - \sigma^2)$, $Z \sim N(0,1)$ and $V \sim \text{Exp}(1)$, with Z and V independent. This can be used to recast Quantile Regression as the following hierarchical model

$$\begin{aligned} Y|X,V &\sim N(X^t\beta + \theta_1\sigma V, \theta_2^2\sigma^2 V) \\ \sigma V &\sim \text{Exp}(\sigma) \end{aligned}$$

which allows for an easy implementation of most MCMC algorithms, see Kozumi and Kobayashi (2011).

1.3.2 Universal Cons of the ALD

Albeit ubiquitous, the ALD has shown numerous drawbacks that may hinder its use in the context of Quantile Regression.

From a computational standpoint, in fact, ALD-based estimation is easy to implement but it is not efficient. The almost exclusive use of MCMC for the Bayesian fitting of Quantile Regression resulted in Bayesian methods for Quantile Regression being slow, but treating the ALD with numerical rather than simulation approaches to optimization has in fact proven to be challenging, as the fact that the function in the exponent is piece-wise linear precludes the use of any solver tailored for smooth functions. Several attempts have thus been made to couple Bayesian Quantile regression with fast fitting procedure by imposing an additional level of smoothing to the check loss in the exponent of the ALD.

For example Yue and Rue (2011) suggests to replace the check loss with

$$\widetilde{\rho}_{\alpha,\gamma}(u) = \begin{cases} \frac{\log(\cosh(\alpha\gamma|u|))}{\gamma} & u \geq 0 \\ \frac{\log(\cosh((1-\alpha)\gamma|u|))}{\gamma} & u < 0 \end{cases}$$

where γ is a fixed parameter such that $\widetilde{\rho}_{\alpha,\gamma}(u) \to \rho_\alpha(u)$ as $\gamma \to \infty$. The value of γ thus tunes the accuracy of the approximation and must be chosen according to the level of the quantile of interest and the amount of data available. Going towards more extreme quantiles will call for a higher value of γ.

Another work in this direction is Fasiolo et al. (2017), which defines a smoother version of the ALD exploiting its connection with the more general family of exponential tails densities defined in Jones (2008) as:

$$p_G(y|\psi,\phi) = K_G^{-1}(\psi,\phi)\exp\left\{\psi y - (\psi+\phi)G^{[2]}(y)\right\}, \tag{1.4}$$

where $\psi, \phi > 0$, $K_G^{-1}(\psi,\phi)$ is a normalizing constant and

$$G^{[2]}(y) = \int_{-\infty}^{y}\int_{-\infty}^{t} g(z)\mathrm{d}z\mathrm{d}t = \int_{-\infty}^{y} G(t)\mathrm{d}t$$

with $g(z)$ and $G(z)$ being the p.d.f. and the c.d.f. respectively of an arbitrary random variable. When $G(z) = \mathbb{I}(z < 0)$, this formulation allows to recover exactly the ALD, hence a smoother version of the ALD can be defined by simply choosing a smoother $G(z)$ function. For example, Fasiolo et al. (2017) suggests to take $G(z)$ to be the c.d.f. of a logistic random variable with scale γ^{-1} and center at 0, i.e.

$$G(z;\gamma) = \frac{\exp\{\gamma z\}}{1+\exp\{\gamma z\}}.$$

As in the previous case, the parameter γ can thought of as the "proximity" to the ALD, as for $\gamma \to \infty$, we have that

$$\frac{\exp\{\gamma z\}}{1+\exp\{\gamma z\}} \to \mathbb{I}[z<0]$$

With this choice of $G(z)$, the density in Eq.1.4 can be rewritten as

$$p_G(y|\alpha,\gamma) = \frac{\gamma \mathrm{e}^{(1-\alpha)y}(1+\mathrm{e}^{y\gamma})^{1/\gamma}}{\mathrm{Beta}\left((1-\alpha)/\gamma, \tau/\gamma\right)}$$

which can be generalized to any location μ and scale σ as

$$p_G(y|\mu,\sigma,\alpha,\gamma) = \frac{\gamma \mathrm{e}^{(1-\alpha)\frac{y-\mu}{\sigma}}(1+\mathrm{e}^{\frac{(y-\mu)\gamma}{\sigma}})^{1/\gamma}}{\sigma \mathrm{Beta}\left((1-\alpha)/\gamma, \alpha/\gamma\right)}. \tag{1.5}$$

Interestingly enough, this density is related to kernel quantile estimation methods. More specifically, by differentiating with respect to μ the log-likelihood corresponding to Eq. (1.5), we obtain

$$\frac{1}{n}\sum_{i=1}^{n} G(y_i;\mu,\sigma,\alpha,\gamma) = 1-\alpha \tag{1.6}$$

where $\frac{1}{n}\sum_{i=1}^{n} G(y_i;\mu,\sigma,\alpha,\gamma)$ is a logistic kernel estimator of the c.d.f., with bandwidth σ/γ, hence the solution of Eq. (1.6) is a standard inversion kernel quantile estimator at $1-\alpha$, as can be seen in Fasiolo et al. (2017). As always in kernel methods, the choice of bandwidth σ/γ is critical. When $\sigma/\gamma \to 0$, the density in Eq. (1.5) converges to the ALD, however, Fasiolo et al. (2017) claims that it is not required nor desirable to approximate the check too closely, as not only it becomes computationally challenging but it is also statistically sub-optimal, in the sense that, as shown in Cheng, Sun, and others (2006) and Falk (1984), kernel estimators are asymptotically better than empirical estimators of the quantiles in terms of relative efficiency.

Both these approaches require the introduction of an additional smoothing parameter, which should be chosen separately for each selected level of the quantile, but they provide a computation-friendly a generalization of the ALD. However, even though its computational drawbacks can be mitigated by means of smoothing, the choice of ALD is still not without consequences, as adopting the ALD imposes several restriction that may not be obvious or desirable in applications. More specifically, as pointed out in Yan and Kottas (2017), assumptions about the data implicitly made in an ALD model are:

- the skewness of the density is fully determined when a specific percentile is chosen (that is, when α is fixed)
- the density is symmetric when $\alpha = 0.5$, that is in the case of median regression
- the mode of the error distribution is at 0 regardless of the parameter α, which results in rigid error density tails for extreme percentiles.

We are not claiming that these assumptions are unreasonable, but rather we are stressing the fact that they should be made if supported by a generating model, and that the ALD assumption does impose some restrictions on the underlying model.

Finally, and more critically for us, from a Bayesian point of view, the major drawback of adopting the ALD as a working likelihood is that posterior inference is restrained by the presence of a confounding parameter. As pointed out by Yang, Wang, and He (2016), the scale parameter σ of the ALD affects the posterior variance, despite not having any impact on the quantile itself. A random variable $X \sim ALD(\alpha, \sigma, \mu)$, in fact, is such that $\mathbb{P}(X \leq \mu) = \alpha$ does not depend on σ (Yue and Rue 2011).

Correction can be made to overcome this issue. For example, Yang, Wang, and He (2016) proposes the following adjustment to the posterior variance to make it invariant with respect to the value of σ:

$$\sqrt{n}\widehat{\Sigma}_{\text{adj}} = \frac{n\alpha(1-\alpha)}{\sigma^2}\widehat{\Sigma}(\sigma)\widehat{D}\widehat{\Sigma}(\sigma)$$

where $\widehat{D} = n^{-1}\sum_{i=1}^{n} X_i X_i^t$ and $\widehat{\Sigma}(\sigma)$ is the posterior variance-covariance matrix. $\widehat{\Sigma}_{\text{adj}}$ however has only asymptotic valid and is limited to model specifications with proper priors. Alternatively, Sriram (2015) proposes a sandwich likelihood method to correct the posterior covariance based on the ALD so that the resulting Bayesian Credible sets satisfy frequentist coverage properties, which is still not a general solution as it tackles the problem from a very specific and partially limited perspective.

1.4 Model–Based Quantile Regression

From a Bayesian perspective, the major drawback of adopting any working likelihood rather than a generating distribution is that the validity of posterior inference is not automatically guaranteed by Bayes Theorem. Our alternative is to reject altogether the use of a working likelihood in favor of the true generating model. We propose a *model–based Quantile Regression*, which exploits the shape of the conditional distribution to link the covariates of interest to the distribution parameter.

Assuming that $Y|X$ is distributed according to some continuous cumulative distribution function $F(y;\theta)$, where θ is the distribution's parameter, our procedure can be formalized in two steps.

- **Modeling step**: the quantile of $Y|X$, $q^\alpha = \mathbb{Q}_\alpha(Y|X)$ is modeled as

  $$q^\alpha = g(\eta^\alpha) \tag{1.7}$$

  where g is an invertible function chosen by the modeler and η^α is the linear predictor, which depends on the level α of the quantile. No restriction is placed on the linear predictor, which can include fixed as well as random effect. Our approach is thus flexible enough to include parametric or semi parametric models, where the interest may lay in assessing the difference in the impact that the covariate may have at different levels of the distribution, as well as fully non parametric models, where the focus is shifted towards prediction instead.

- **Mapping step**: the quantile q^α is mapped to the parameter θ as

  $$\theta = h(q^\alpha, \alpha) \tag{1.8}$$

> where the function h must be invertible to ensure the identifiability of the model and explicitly depends on the quantile level α. The map h gives us a first interpretation of model-based Quantile Regression as a reparametrization of the generating likelihood function $F(y;\theta)$, in terms of its quantiles, i.e. $F(y; q^\alpha = h^{-1}(\theta, \alpha))$.

When $\theta \in \mathbb{R}$, the map $\theta = h(q^\alpha, \alpha)$ is uniquely defined. When $\theta \in \mathbb{R}^d$, with $d > 1$, all the components of the model parameters have to be redefined as a function of the quantiles. As the quantile is a location parameter, linking the location parameter of the model to it directly it is straightforward. For the other parameters of the distribution, there are multiple option to be explored; depending on the meaning of the parameter one could rewrite as a function of interquartile distance, which represent a measure of variability, or exploiting Groeneveld and Meeden's coefficient for skewness (Groeneveld and Meeden 1984).

By linking the quantiles of the generating distribution to its canonical parameter θ, we are indirectly modeling θ as well, hence we are implicitly building a connection between Quantile Regression and Generalized Linear (Mixed) Models (GLMM), which are also concerned with the modeling of θ. The modeling and mapping steps in fact, can be considered as a way to define a link function, in the GLMM sense, as the composition $\theta = h(g(\eta))$, and this allows us to rephrase Quantile Regression as a new link function in a standard GLMM problem. Drawing a path from GLMM to Quantile Regression is instrumental in the fitting however the pairing is only formal: coefficients and random effect have a completely different interpretation.

One of the advantages of coupling Quantile Regression to GLMM is that it allows to bypass slow MCMC methods for the fitting and instead use `R-INLA` (Rue et al. 2017), which allows for both flexibility in the model definition and efficiency in their fitting. `R-INLA` is an `R` package that implements the `INLA` (Integrated Nested Laplace Approximation, see Rue, Martino, and Chopin (2009)) method for approximating marginal posterior distributions for hierarchical Bayesian models, where the latent structure is assumed to be a Gaussian Markov Random Field, (Rue and Held 2005), which is especially convenient for computations due to its naturally sparse representation. The class of model that can be formulated in a hierarchical fashion is broad enough to include parametric and semi parametric models; the `INLA` approach is thus extremely flexible and provides a unified fitting framework for mixed and additive models in all their derivations. More details are provided in Appendix A.

1.5 Discrete Data

Quantile regression was originally defined for continuous responses and extending it to the case of discrete variable has proven to be challenging.

In the "classical" model-free setting, inference is limited by the fact that the non-differentiable objective function in Equation (1.2) together with the points of positive mass of one of the variables involved in the optimization problem, makes it impossible to derive an asymptotic distribution for the sample quantiles (Machado and Santos Silva 2005).

In the case of model-based Quantile Regression, dealing with discrete distribution is non trivial since it is difficult to define both the model g and the map h as defined in Section

1.4. As far as the modeling step is concerned, most common model choices, such as the log model for count data or logit for binary responses are typically continuous and are not well suited to represent the conditional quantiles, which are intrinsically discrete. More tragically, as the quantile function is discrete, it is not possible to define an injective map h, which means that it is not possible to define a unique θ generating each quantile, as can be seen in Figure 1.5.

Either way, in order to fit the model it is necessary to impose some additional level of smoothing, which is usually done by approximating the distribution of the discrete random variable with a continuous analogue.

1.5.1 Jittering

In order to enforce the necessary level of smoothing, the most natural approach is to treat the discrete responses as if they were generated by a continuous distribution. This is a common element to most strategies for dealing with quantiles for discrete data, which then differ in how this continuous distribution is built/chosen.

To date, the most famous strategy to rephrase Quantile Regression for discrete data in a continuous setting is *jittering*, first introduced by Machado and Santos Silva (2005), which consists in adding continuous and bounded noise U to the response variable Y and then model quantiles of the "continuous" $Z = Y + U$. The noise variable U is typically taken to be Uniformly distributed in $(0, 1)$, although the procedure would hold for any bounded continuous distribution, as long as U and Y are independent.

The quantiles of the new random variable Z are in one-to-one relation with the quantiles of the original variable of interest Y, in the sense that $Q_Y(\alpha) = \lceil Q_Z(\alpha) - 1\rceil$. However, although the distribution of the new random variable Z is continuous, it is not smooth over the entire support, since it does not have continuous derivative for integer values of Z, hence additional assumptions are needed in order to carry out inferential procedures.

Moreover, the estimates $\hat{\beta}_\alpha$ are naturally affected by the specific realization of the jittering noise, hence it is advisable to remove its effect by means of averaging or integrating. In the first case, Machado and Santos Silva (2005) defines the *average-jittering* estimator $\hat{\beta}_{m,\alpha}$ as

$$\hat{\beta}_{m,\alpha} = \frac{1}{m} \sum_{j=1}^{m} \hat{\beta}_\alpha^{(j)}$$

where $\hat{\beta}_\alpha^{(j)}$ with $j = 1, \ldots, m$ is any estimate of β_α computed on the j-th jittered sample $\mathcal{D}_n^{(j)} = \{Z_i = Y_i + u_i^{(j)}, X_i\}_{i=1}^n$, and $u^{(1)}, \ldots, u^{(m)}$ are independent random samples.

1.5.2 Model-aware Interpolation

Jittering can be thought of as a way of interpolating a discrete distribution that it is insensitive to the specific features of the distribution. In order to tailor the interpolation on the model we are considering, we define continuous distributions to use in the fitting that are aware of the shape of the original discrete distribution. Inspired by Ilienko (2013), we focus in particular on discrete distributions whose c.d.f. can be written as

$$F_X(x; \theta) = \mathbb{P}(X \leq x) = k(\lfloor x \rfloor, \theta) \tag{1.9}$$

where k is a continuous function in the first argument. The continuous interpolation is then defined by removing the floor operator, so that the function $k(x, \theta)$ is the c.d.f. of X', a continuous version of X. By definition of floor we have that

$$F_X(x) = k(\lfloor x \rfloor, \theta) = k(x, \theta) = F_{X'}(x) \tag{1.10}$$

for all integer x, and since the two c.d.f.s are the same at the integer values of x, this can be seen as a continuous generalization of the original variable.

One of the features of our model-aware strategy is that it allows for a proper assessment of variability. In Section 1.4 we already made a case for the advantages of the model-based approach in the Bayesian setting, however, this is profitable in the frequentist framework as well. Confidence intervals for the regression coefficient in fact heavily rely on the asymptotic normality of the sample quantiles, which is guaranteed only when the distribution generating the data is continuous, while sample quantiles of discrete distribution in general are not asymptotically normal. It is possible to generalize the definition of sample quantile to quantile of the mid–distribution in order to gain asymptotic normality to exploit in the construction of confidence intervals in the discrete setting (Ma, Genton, and Parzen 2011), however, the interpretation of this new summary is not entirely clear.

Even in the ubiquitous jittering approach, it is cumbersome to determine the variance associated to the estimates for the coefficients; the asymptotic normality of the sample quantiles is granted in fact as the sample size as well as the number of repetition of the jittering procedure go to infinity, and thus variance estimation require computationally intensive re-sampling procedures.

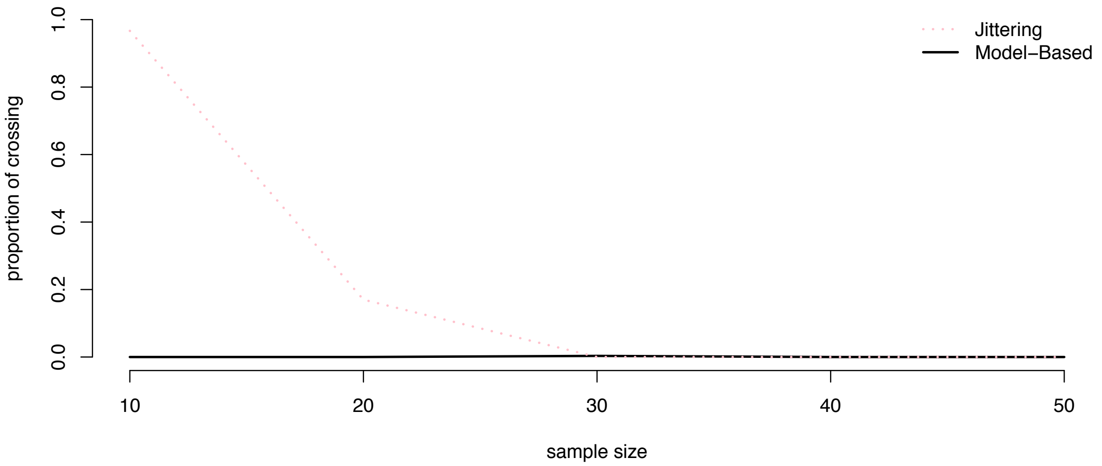

Figure 1.3: Estimated probability of having at least one crossing on a simulated example where $Y|X \sim \text{Poisson}(\exp\{X\})$, where $X \sim \text{Gamma}(1, 5)$.

Moreover, our approach has shown to be less prone than jittering to the phenomenon of *quantile crossing*. This is a paradoxical result occurring when the quantile curves intersect one another, as shown in Figure 1.4, resulting in the total loss of meaning of the curves themselves. Despite distributional quantiles being by definition an increasing function of the probability index, this is not always true for the fitted curves, especially when the sample size is small.

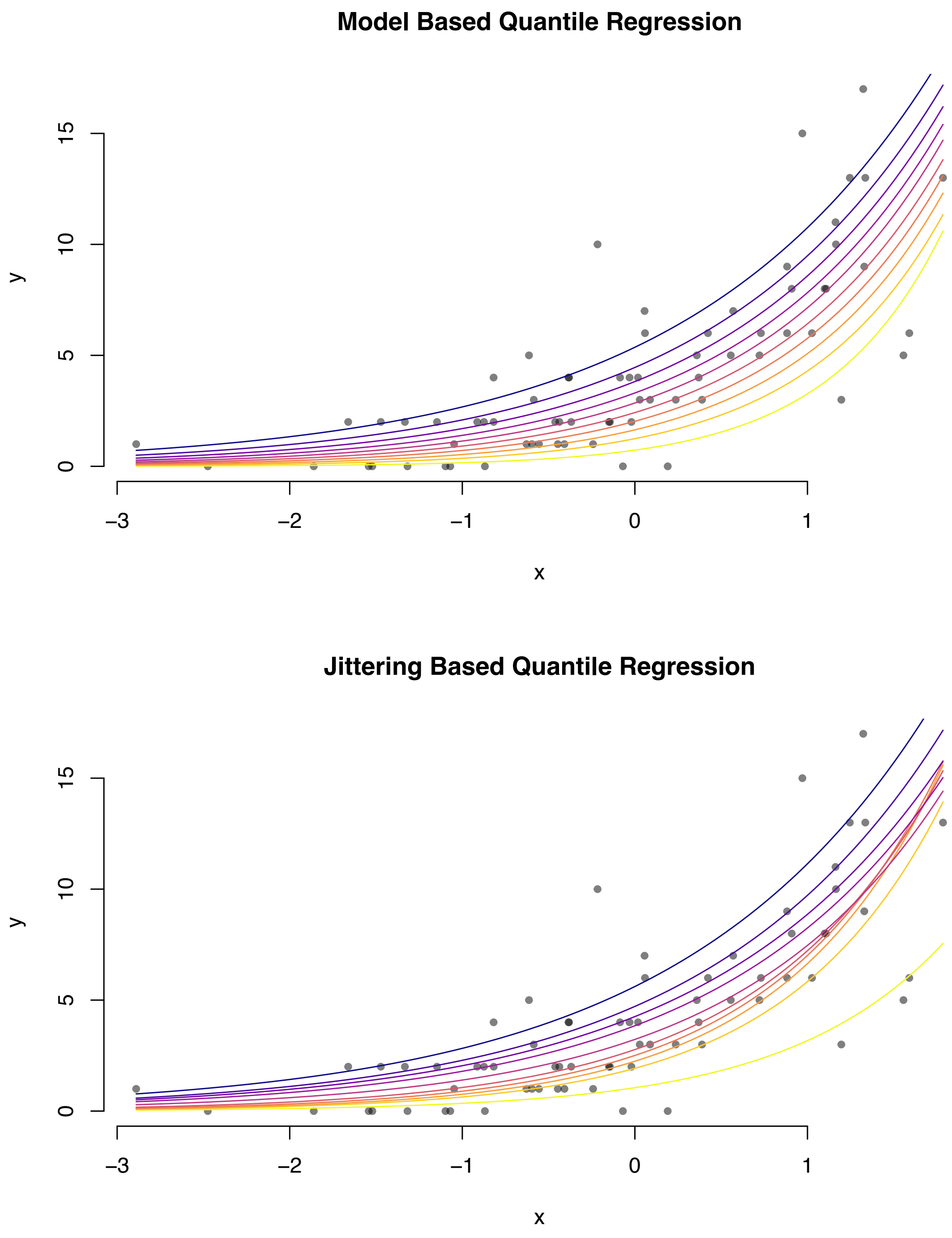

Figure 1.4: Quantile curves estimated with model based Quantile Regression (top) and jittering (bottom). Darker curves correspond to higher quantile levels.

Figure 1.3 shows how, even on a simulated toy example, the model based quantile estimator seems to be less affected by crossing.

1.6 Continuous Poisson Distribution

The class of distribution defined by Equations (1.9) and (1.10) is broad enough to include the three distribution most frequently encountered in applications: Poisson, Binomial and Negative Binomial. We explore in detail the Poisson case, the other two are trivial extensions.

The starting point for defining a continuous version of the Poisson distribution is to rewrite the cumulative density function for the Poisson as the ratio of Incomplete and Regular Gamma function:

$$X \sim \text{Poisson}(\lambda) \qquad F_X(x) = \mathbb{P}(X \leq x) = \frac{\Gamma(\lfloor x \rfloor + 1, \lambda)}{\Gamma(\lfloor x \rfloor + 1)} \qquad x \geq 0 \tag{1.11}$$

where

$$\Gamma(x, \lambda) = \int_{\lambda}^{\infty} e^{-s} s^{x-1} \mathrm{d}s$$

is the upper incomplete Gamma function. Extending the Poisson distribution to the continuous case from this formulation is just a matter of removing the floor operator, that is

$$X' \sim \text{Continuous Poisson}(\lambda) \qquad F_{X'}(x) = \mathbb{P}(X' \leq x) = \frac{\Gamma(x+1, \lambda)}{\Gamma(x+1)} \qquad x > -1. \tag{1.12}$$

where the domain has been extended from $x \geq 0$ to $x > -1$ in order to avoid mass at 0.

The Continuous Poisson defined in Equation (1.12) is similar to that of Ilienko (2013), with the noticeable difference that our definition of Continuous Poisson is shifted by 1, so that the Discrete Poisson X is a monotonic left continuous function of the Continuous Poisson X'; more specifically Continuous and Discrete versions of the Poisson are related by

$$X = \lceil X' \rceil. \tag{1.13}$$

By integration by parts we have

$$\frac{\Gamma(x+1, \lambda)}{\Gamma(x+1)} - \frac{\Gamma(x, \lambda)}{\Gamma(x)} = \lambda^x e^{\lambda} / \Gamma(x+1) \tag{1.14}$$

which is enough to show that:

$$\begin{aligned}
\mathbb{P}(\lceil X' \rceil = x) &= \mathbb{P}(X' \in (x-1, x]) = F_{X'}(x) - F_{X'}(x-1) \\
&= \frac{\Gamma(x+1, \lambda)}{\Gamma(x+1)} - \frac{\Gamma(x, \lambda)}{\Gamma(x)} = \lambda^x e^{\lambda} / \Gamma(x+1) \\
&= \mathbb{P}(X = x). \qquad (1.15) \\
& \qquad (1.16)
\end{aligned}$$

Following Ilienko (2013), we have that $F_{X'}(x)$ is a well defined c.d.f., in the sense that it is non-decreasing in x, is right-continuous and it satisfies:

$$\lim_{x \to \infty} F_{X'}(x) = 1 \qquad \text{and} \qquad \lim_{x \to -\infty} F_{X'}(x) = 0.$$

The good thing about this definition is that $F_X(k) = F_{X'}(k) \quad \forall\, k \in \mathbb{N}$. The bad thing about this definition is that it is not yet completely continuous, as there is a jump in $x = 0$. In order to overcome this, we thus change our definition to

$$X' \sim \text{Continuous Poisson}(\lambda) \qquad\qquad F_{X'}(x) = \frac{\Gamma(x+1,\lambda)}{\Gamma(x+1)} \times \mathbb{1}\{x \geq -1\} = G_\lambda(x+1).$$

where $G_\lambda(x)$ is known as regularized upper incomplete gamma function. As can be seen in Figure 1.5, this new distribution is still an interpolant of the original discrete one. Notice that although it is "technically" allowed for $x < 0$, as we are using this distribution to model counts, this is not necessary nor used for our purposes. It is also worth noticing that $1 - G_\lambda(x)$ corresponds to the c.d.f. of a Gamma distribution with parameters $(x, 1)$, i.e.

$$G_\lambda(x+1) = 1 - F_{\text{Gamma}(x+1,1)}(\lambda).$$

It is possible to determine a density function for the Continuous Poisson distribution, which looks like

$$\begin{aligned} f_X(x;\lambda) &= \frac{G^{3,0}_{2,3}(|\lambda)}{\Gamma(x)} + Q(x,\lambda)\left[\log(\lambda) - \psi^{(0)}(x)\right] \\ &= \frac{\lambda^x_2 G^{3,0}_{2,3}(|\lambda)}{\Gamma(x)} + P(x,\lambda)\left[\psi^{(0)}(x) - \log(\lambda)\right] \end{aligned}$$

where G is the Meijer G-function and F is the generalized hypergeometric function, $\psi(k)(x)$ is the k-th derivative of the digamma function. However this is not easy to work with and implies that the moments of the Continuous Poisson do not have a closed form expression but they require numerical approximations.

Interestingly enough, the moments of the continuous and discrete Poisson distribution do not coincide, although asymptotically their ratio tends to 1.

1.6.1 Continuous Count distributions

The Binomial and the Negative Binomial distribution can also be trivially extended to the continuous case. Their c.d.f. can in fact be written as:

$$Y \sim \text{Binomial}(n,p) \qquad F_Y(y) = I_{1-p}(n - \lfloor y \rfloor, \lfloor y \rfloor + 1) \tag{1.17}$$

$$Z \sim \text{Negative Binomial}(r,p) \qquad F_Z(z) = I_{1-p}(r, \lfloor z \rfloor + 1) \tag{1.18}$$

where $I_x(a,b)$ is the *regularized incomplete Beta function* defined as:

$$I_x(a,b) = \frac{\mathrm{B}(a,b,x)}{\mathrm{B}(a,b)} \qquad \text{with} \qquad \mathrm{B}(a,b,x) = \int_x^1 t^a (1-t)^{b-1} \mathrm{d}t \tag{1.19}$$

Again the extension of these two random variables to the continuous case can be obtained by removing the floor operator:

$$Y' \sim \text{Continuous Binomial}(n,p) \qquad F_{Y'}(y) = I_{1-p}(n - y, y + 1) \tag{1.20}$$

$$Z' \sim \text{Continuous Negative Binomial}(r,p) \qquad F_{Z'}(z) = I_{1-p}(r, z + 1) \tag{1.21}$$

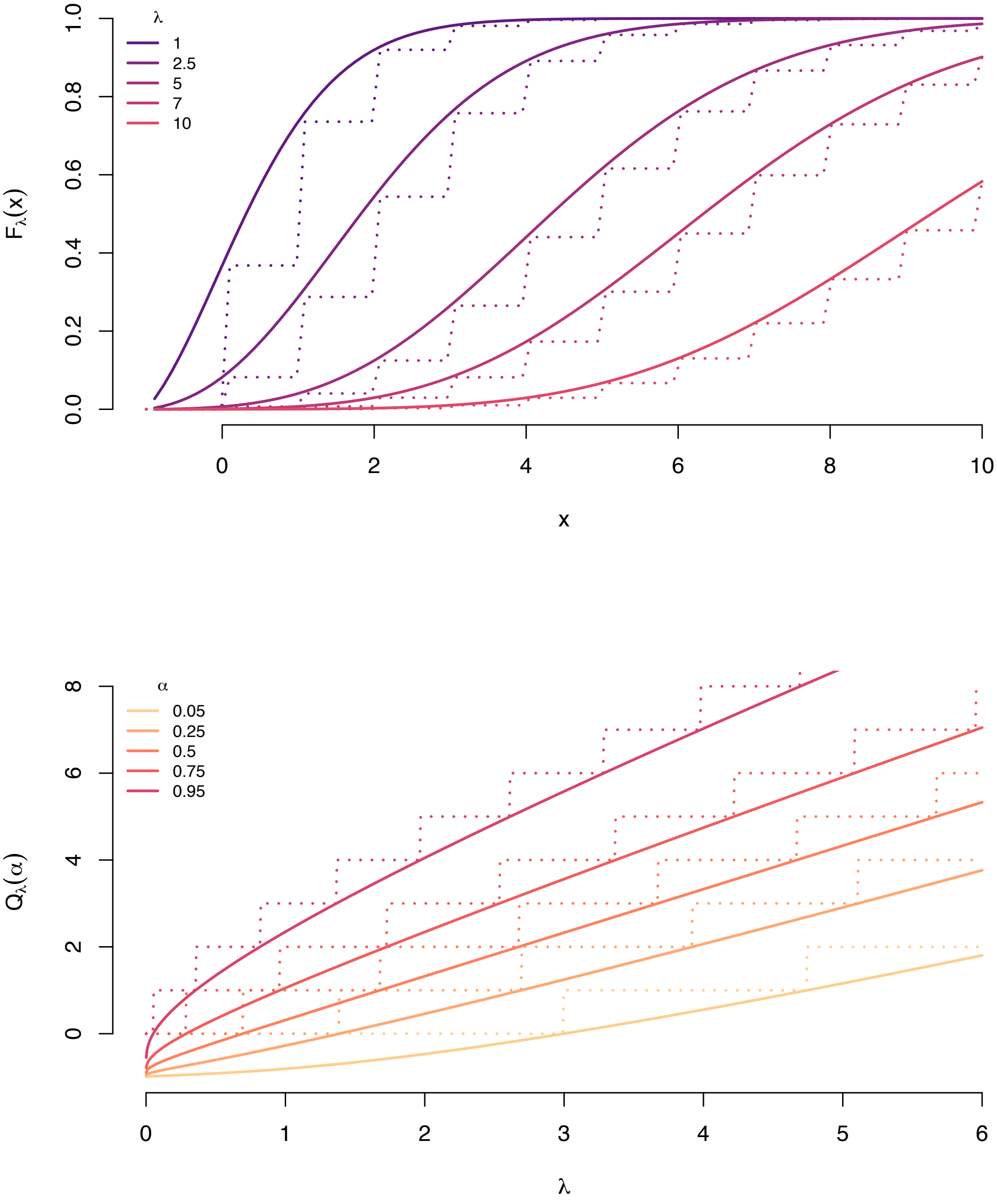

Figure 1.5: Top: c.d.f. of the discrete (dashed line) and continuous (continuous line) Poisson distribution for several values of λ. Bottom: quantile function of the discrete (dashed line) and continuous (continuous line) Poisson distribution.

It is obvious that these two continuous distributions, exactly like the continuous Poisson defined before, result in interpolation of both the c.d.f. and quantile functions of the discrete original, analogously to what can be seen in Figure 1.5.

We previously claimed that the main advantage of this model-aware strategy for approximating discrete distributions with continuous versions, it is that the new continuous variables retain the same structure of their discrete counterpart. This can be made more explicit in the case of the Poisson, Binomial and Negative Binomial, where the behavior of the resulting continuous random variables mimic that of their discrete counterparts.

In the discrete case in fact it is well known that the Poisson distribution is the limiting case of both the Binomial and the Negative Binomial when the probability of observing one event goes to 0 and that Binomial and Negative Binomial are also entwined in a 1-to-1 relation. The same relations are preserved in the continuous case, hence the two classes of distribution have similar meaning.

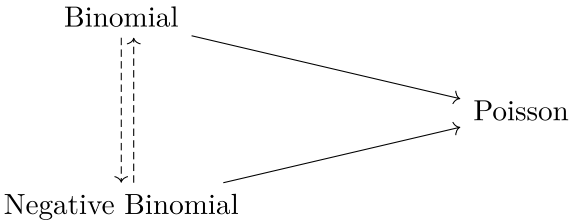

Figure 1.6: Diagram summarizing the connection between Binomial, Negative Binomial distribution. Continuous lines indicate asymptotic relations, dashed lines denote a finite sample relation.

Poisson and Binomial Let X be a Continuous Poisson with parameter λ, Y be a Continuous Binomial with parameters n and p. Then by following Ilienko (2013) we have that for $n \to \infty$ and $p \to 0$ so that $np \to \lambda$

$$F_Y(x) = \frac{\mathrm{B}(x+1, N-x, p)}{\mathrm{B}(x+1, N-x)} \longrightarrow \frac{\Gamma(x+1, \lambda)}{\Gamma(x+1)} = F_X(x). \tag{1.22}$$

Analogously to its discrete version, the Continuous Poisson can thus be interpreted as an approximation for a binomial-like distribution in the case of rare events.

Poisson and Negative Binomial Let X be a Continuous Poisson with parameter λ, Z be a Continuous Negative Binomial with parameters r and p. Then it follows trivially from Equation (1.22) that for $r \to \infty$ and $p \to 0$ so that $rp \to \lambda$ we have

$$F_Z(x) = \frac{\mathrm{B}(x+1, r, p)}{\mathrm{B}(x+1, r)} \longrightarrow \frac{\Gamma(x+1, \lambda)}{\Gamma(x+1)} = F_X(x). \tag{1.23}$$

From a modeling perspective, this motivates the choice of the Continuous Negative Binomial instead of the Continuous Poisson in cases where there is over-dispersion, i.e. the assumption of mean and variance being equal is clearly violated.

Binomial and Negative Binomial Let Z be a Continuous Negative Binomial with parameters r and p and Y be a Continuous Binomial with parameters $s+r$ and $1-p$, then

$$\begin{aligned}F_Z(s) &= 1 - I_p(s+1, r)\\ &= 1 - I_p((s+r) - (r-1), (r-1)+1)\\ &= 1 - \mathbb{P}(Y \leq r-1)\\ &= \mathbb{P}(Y \geq r)\end{aligned} \tag{1.24}$$

which justifies the interpretation of the Continuous Negative Binomial as the waiting time until the arrival of the r-th success in a Binomial-like experiment.

1.7 Quantile Regression for Poisson data

By assuming that discrete responses are generated by a Continuous Poisson discrete, it is possible to extend Quantile Regression to count data. If $Y|\eta \sim$ Continuous Poisson(λ), in fact, we can specify the link function g and the parameter map h. More specifically we have:

$$g: \qquad q^\alpha = g(\eta^\alpha) = \exp\{\eta^\alpha\} \tag{1.25}$$

$$h: \qquad \lambda = h(q^\alpha) = \frac{\Gamma^{-1}(q^\alpha+1, 1-\alpha)}{\Gamma(q^\alpha+1)}. \tag{1.26}$$

The Poisson distribution is typically used to model count data, however it is also possible to use it when modelling rates. This is especially useful in the context of aggregated data, such counts over a time interval or counts over an area, where the aggregating variable (length of the time interval, or the size of the geographical area for example) affects the distribution and makes comparison between units meaningless. When units are subjected to different exposures E, there are two ways of encoding it into the model:

- by including them in the model as offset, discounting the quantiles directly and considering q^α/E

$$\begin{aligned}q^\alpha &= \exp\{\eta^\alpha + \log(E)\} = E\exp\{\eta^\alpha\}\\ \lambda &= \frac{\Gamma^{-1}(q^\alpha+1, 1-\alpha)}{\Gamma(q^\alpha+1)}\end{aligned} \tag{1.27}$$

- by adjusting the global parameter and consider λ/E

$$\begin{aligned}q^\alpha &= \exp\{\eta^\alpha\}\\ \lambda &= E\frac{\Gamma^{-1}(q^\alpha+1, 1-\alpha)}{\Gamma(q^\alpha+1)}.\end{aligned} \tag{1.28}$$

While in Poisson mean regression these two approaches yield the same results, as

$$\lambda = E\exp\{\eta\} \iff \lambda/E = \exp\eta \tag{1.29}$$

in Poisson Quantile Regression this is not true. In general

$$\frac{\Gamma^{-1}(E\exp\{\eta_i^\alpha\}^\alpha + 1, 1-\alpha)}{\Gamma(E\exp\{\eta_i^\alpha\}^\alpha + 1)} \neq E\frac{\Gamma^{-1}(\exp\{\eta_i^\alpha\}^\alpha + 1, 1-\alpha)}{\Gamma(\exp\{\eta_i^\alpha\}^\alpha + 1)} \tag{1.30}$$

and, besides the trivial case $E = 1$, it is not obvious to determine whether there are values of E for which the equality would hold since there is no closed form solution for Γ^{-1}. A case could be made for both modeling strategies, the former being a "quantile-specific" model while the latter being more of a global model, and choosing between them depends on the application.

As can be seen in Figure 1.5, the quantiles of the two distributions are not the same, and the regression model returns fitted quantiles for the Continuous Poisson. However, fitted quantiles of the discrete distribution can be obtained by exploiting quantile equivariance, since we defined the continuous Poisson so that its discrete counterpart is a monotonic left continuous function. Let $Y|\eta \sim$ Poisson(λ) and $Y'|\eta \sim$ Continuous Poisson(λ), then we have

$$\mathbb{Q}_\alpha(Y'|\eta) = \mathbb{Q}_\alpha(\lceil Y \rceil|\eta) = \lceil \mathbb{Q}_\alpha(Y|\eta) \rceil. \tag{1.31}$$

1.7.1 An application - Disease Mapping

We conclude by showing with an application the potential of Quantile Regression in the context of disease mapping, when the goal of the analysis to identify which areas correspond to a high risk. We show this using emergency hospitalization data as in Congdon (2017).

Our dataset consist of:

- $\texttt{Y}$: counts of emergency hospitalizations for self-harm collected in England over a period of 5 years (from 2010 to 2015). The counts are aggregated over 6791 are Middle Level Super Output Areas (MSOAs).
- $\texttt{X}_1$: Deprivation, as measured by the 2015 Index of Multiple Deprivation (IMD).
- $\texttt{X}_2$: Social fragmentation, measured by a composite index derived from indicators from the 2011 UK Census comprising housing condition and marital status.
- $\texttt{X}_3$: Rural status, again measured by a composite indicator aimed at capturing the accessibility to services and facilities such as schools, doctors or public offices.

Standard risk measure, such as the ratio between observed and expected cases in each area, the Standardized Mortality (or Morbidity) Ratio (SMR) $SMR = \texttt{Y}/\texttt{E}$, is not reliable here due to the high variability of expected cases $\texttt{E}$ (Figure 1.8), hence is advisable to introduce a random effect model that exploit the spatial structure to obtain more stable estimates of the risk. Assuming that, conditionally on covariates $\texttt{X}$ and random effects b, the observations are generated by a Poisson distribution

$$\texttt{Y}|\texttt{X}, b \sim \text{Poisson}(\lambda) \tag{1.32}$$

we adopt the following model for the conditional quantile of level α

$$\mathbb{Q}_\alpha(\texttt{Y}|\texttt{X}, b) = \texttt{E}\theta_\alpha = \texttt{E}\exp\{\eta\}. \tag{1.33}$$

We opted for the *quantile-level* approach for handling exposures $\texttt{E}$ in order to ease interpretation; as we discount each quantile for the exposures, in fact, the parameter $\theta_{i,\alpha}$ corresponding to the $i^{\texttt{th}}$ area can be considered the relative risk of unit i at level α of the population. The linear predictor η can be decomposed into

$$\eta = \beta_0 + \beta_{\texttt{Depr}}\texttt{X}_1 + \beta_{\texttt{SF}}\texttt{X}_2 + \beta_{\texttt{RS}}\texttt{X}_3 + b \tag{1.34}$$

where β_0 represent the overall risk and b consists in the sum of an unstructured random effect capturing overdispersion and measurement errors and spatially structured random effect. In order to avoid the confounding between the two components of the random effect and to avoid scaling issues we adopt for b the modified version of the Besag–York–Mollier (BYM) model introduced in Simpson et al. (2017):

$$b = \frac{1}{\tau_b}\left(\sqrt{1-\phi}v + \sqrt{\phi}u\right). \tag{1.35}$$

Both random effects are normally distributed, and in particular

$$v \sim N(0, \mathbf{I}) \tag{1.36}$$

$$u \sim N(0, Q_u^{-1}) \tag{1.37}$$

so that $b \sim N(0, Q_b^{-1})$ with $Q_b^{-1} = \tau_b^{-1}(1-\phi)\mathbf{I} + \phi Q_u^{-1}$, a weighted sum of the precision matrix for the $\mathbf{I}$ and the precision matrix representing the spatial structure Q_u, scaled in the sense of Sørbye and Rue (2014).

We assign priors on the precision τ_b and the mixing parameter ϕ using the penalized complexity (PC) approach, as defined in Simpson et al. (2017) and detailed in Riebler et al. (2016) in the special case of disease mapping. Estimated coefficients shown in Table 1.3 show that Deprivation has a negative impact, which only slightly attenuates at higher quantile level, meaning that, as we could expect, higher deprivation corresponds to increases in self harm hospitalization. Interestingly, being a rural area seems to have a positive effect instead, with more rural areas being associated to lower rates of hospitalization.

Despite regression being a key tool for disease mapping, the use of Quantile Regression instead of mean regression is still unexplored, with exceptions in Congdon (2017) and Chambers, Dreassi, and Salvati (2014). This is somehow surprising, since the focus of disease mapping is on extreme behaviors of the population, for which using quantiles, that provide insights on the tails of the distributions, would seem a more natural choice than considering means. The relative risk $\theta_{i,\alpha}$ can be directly used to detect "high risk" areas. Following Congdon (2017), the $i^{\mathtt{th}}$ area region is considered at "high risk" if $[\theta_{i,0.05}, \theta_{i,0.95}] > 1$, where 1 represents an increase in the risk, otherwise it is assumed to be "low risk".

	Mean	1st Quartile	2nd Quartile	3rd Quartile
β_0	-0.5989 (0.0151)	-0.7076 (0.2650)	-0.4701 (0.2758)	-0.5128 (0.0420)
$\beta_{\mathtt{Depr}}$	1.9810 (0.0312)	2.0871 (0.2202)	1.9608 (0.1570)	1.9340 (0.0599)
$\beta_{\mathtt{RS}}$	-0.8148 (0.0364)	-0.8834 (0.1270)	-0.8781 (0.2170)	-0.7820 (0.2170)
$\beta_{\mathtt{SF}}$	0.4291 (0.0447)	0.5628 (0.1558)	-0.0981 (0.8993)	0.3997 (0.1052)
τ_b	6.4098 (0.1996)	5.7681 (0.1807)	6.2743 (0.1960)	7.1589 (0.1772)
ϕ	0.8386 (0.0143)	0.8383 (0.0144)	0.8387 (0.0143)	0.8172 (0.0116)

Table 1.3: Posterior mean estimates of model parameters (and corresponding standard deviations).

Mean regression methods for identifying "high risk" areas are also based on relative risk θ, although defined in a different way, i.e.

$$\mathbb{E}[\mathtt{Y}|\mathtt{X}, b] = \mathtt{E}\theta = \mathtt{E}\exp\{\eta\}. \tag{1.38}$$

Posterior probability of an increase in the risk are then used to assess whether an area has high risk or not, so that the $i^{\mathtt{th}}$ area is considered to be of high risk if $\mathbb{P}(\theta_i > 1|Y_1, \dots Y_n) > t$ where t is a threshold value depending on the application (in this case we chose $t = 0.9$).

The difference between the two methods is that in the former high risk areas are those where the risk increases for every level of the population, i.e. for those areas which are very sensible and those which are less sensible to the disease, while the latter considers only the mean level, which is a synthetic measure for the whole population but it may be subject to compensation. Figure 1.9 and 1.10 show the critical areas identified by quantile and mean regression. The similarity of the results of our method with those corresponding to a more traditional approach, as well as to previous analyses, reassures us that our method yields reasonable results. At the same time, the minor discrepancies between the two maps is also encouraging, as the two methods have different definitions of high risk; different results correspond in fact to different insights on the disease risk and the non-overlap between quantile-based exceedance probability-based methods testifies that there is information to be gained from our approach.

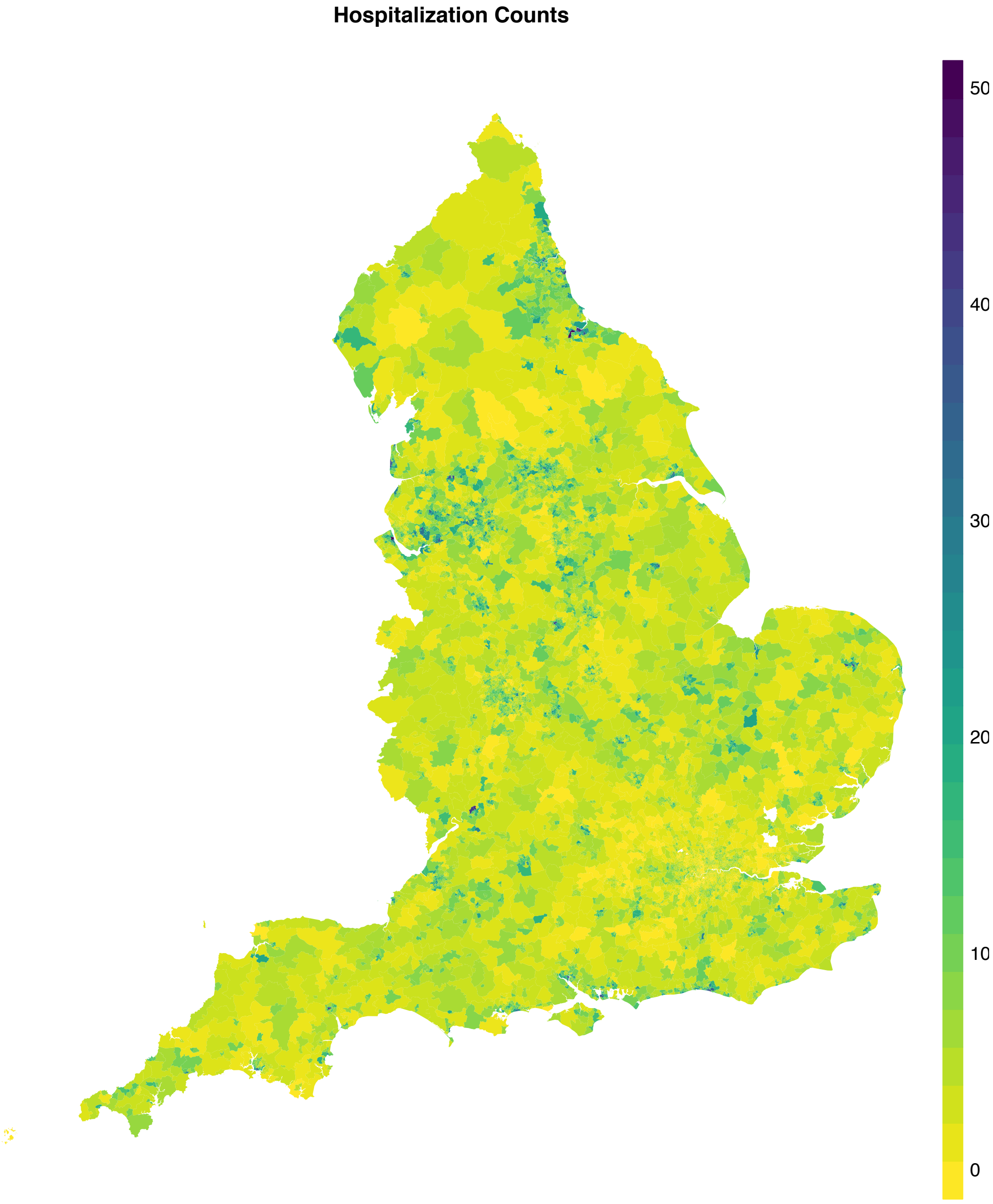

Figure 1.7: Raw counts of self harm hospitalization for the MSOAS.

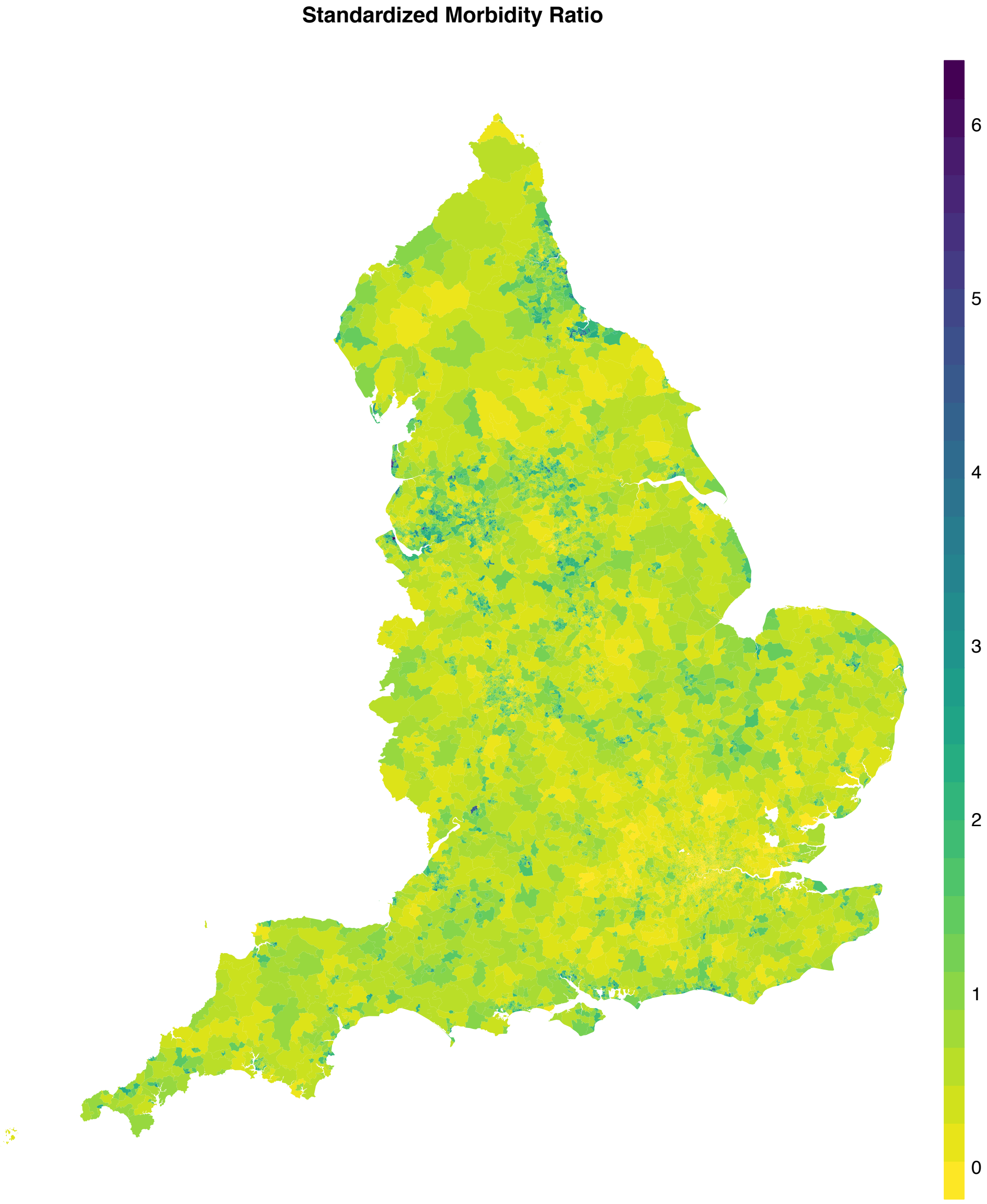

Figure 1.8: Quantile Relative Risk of level 0.025. In gray areas of High Risk.

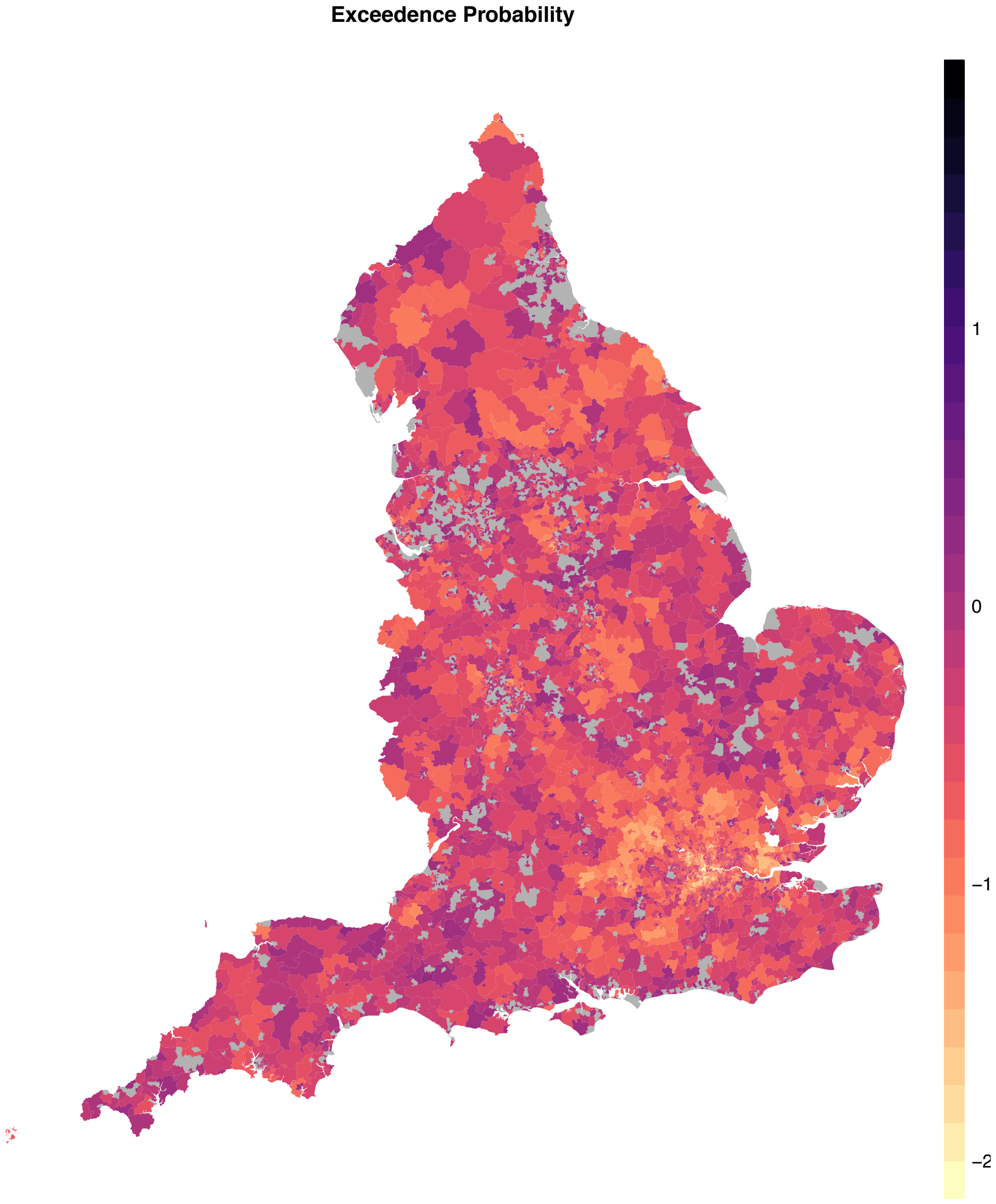

Figure 1.9: Exceedence probability for Mean Relative Risk. In gray areas of High Risk.

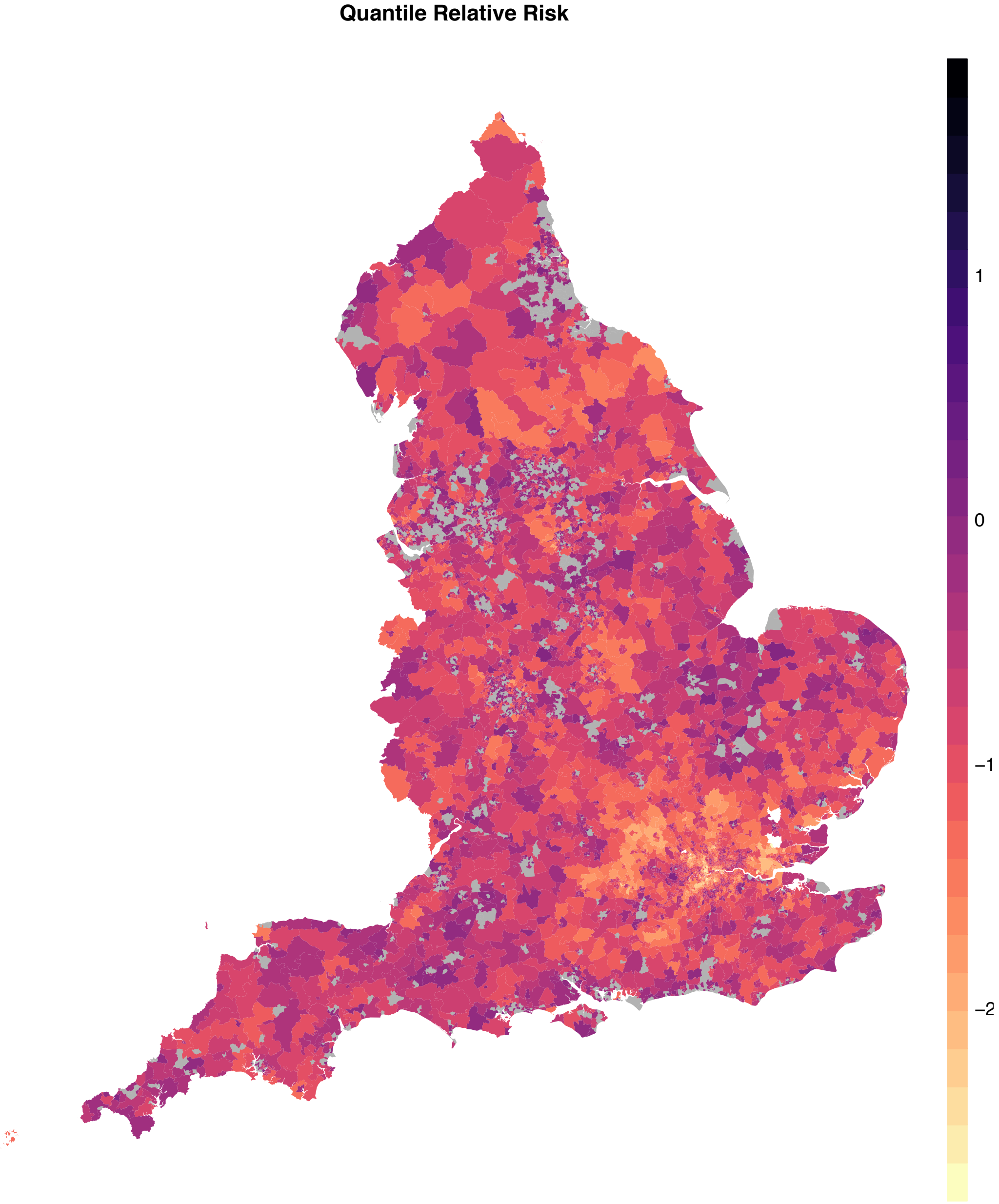

Figure 1.10: Quantile Relative Risk of level 0.025. In gray areas of High Risk.

Chapter 2

Topological Tools for Data Analysis

2.1 The shape of fixed-scale data

As we are dealing with increasingly complex data, our need for characterizing them through a few, interpretable features has grown considerably. Topology has proven to be a useful tool in this quest for "insights on the data", since it characterizes objects through their connectivity structure, i.e. connected components, loops and voids. In a statistical framework, this characterization yields relevant information: for example, connected components correspond to clusters (Chazal et al. 2013) while loops represent periodic structures (Perea and Harer 2015). At the crossroad between Computational Topology and Statistics, Topological Data Analysis (`TDA` from here onwards) is a new and expanding research area devoted to recovering the shape of the data focusing in particular on its topological structure (Carlsson 2009).

Although topology has always been considered a very abstract branch of mathematics, it has some properties that are extremely desirable in data analysis, such as:

- *It does not depend on the coordinates of the data, but only on pairwise distances.* In many applications, coordinates are not given to us or, even if they are, they have no meaning and they could be misleading.
- *It is invariant with respect to a large class of deformations.* Two object that can be deformed into one another without cutting or gluing are topologically equivalent, meaning that topological methods are flexible.
- *It allows for a discrete representation of the objects we study.* Most continuous objects can be approximated with a discrete but topologically equivalent object, for which it is easier to define algorithms.

2.1.1 Persistent Homology Groups - Intuition

The broad goal of `TDA` is to recover the topological structure (i.e. 0-dimensional topological features or connected components, 1-dimensional topological features or cycles and so on) of any arbitrary function of data f, by characterizing it in terms of some topological invariant, most often its Homology Groups, while also providing a measure of their importance.

The main advantage of this choice in terms of interpretability is that Homology Groups of dimension k represent k-dimensional connected structures: the Homology Group of dimension 0 of a topological space $\mathbb{X}$, $H_0(\mathbb{X})$ represents connected components of $\mathbb{X}$, the Homology Group of dimension 1, $H_1(\mathbb{X})$ represents loops (or cycles) of $\mathbb{X}$, $H_2(\mathbb{X})$ represents voids, and so on (we refer to Appendix B for a brief introduction of Homology Groups or to Hatcher (2002) for a more complete and rigorous treatment of the subject).

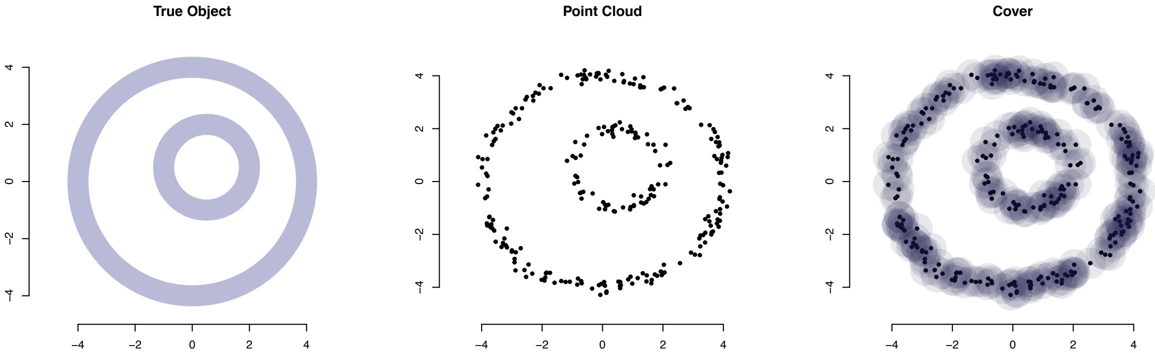

Figure 2.1: From left to right: the true object we are trying to recover ($\mathbb{X}$), the point–cloud data sampled on it ($\mathbb{X}_n$) and the cover (d_ε).

In practice we most often do not observe the object we are interested in $\mathbb{X}$ directly, but a point–cloud $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ sampled on it, which may not be explored using Homology Groups directly. A point–cloud $\mathbb{X}_n$, in fact, has a trivial topological structure per se, as it is composed of as many connected components as there are observations and no higher dimensional features. Topological Invariants in the TDA framework are thus built from functions of $\mathbb{X}_n$ rather than on the point–cloud itself, using an extension of Homology, *Persistent Homology*, which is the mathematical backbone of TDA (Edelsbrunner and Harer 2010).

Roughly speaking, Persistent Homology provides a characterization of the topological structure of any arbitrary function f by building a filtration on it (typically its sublevel or superlevel sets, f_ε and f^ε respectively). The link between Persistent Homology and the "shape of the data" is that for some choice of f, sublevel (or respectively superlevel) set filtrations are topologically equivalent to the space data was sampled from, $\mathbb{X}$.

Distance functions The most common choice for analysing the topological structure of $\mathbb{X}$ is to investigate the Persistent Homology of the sublevel-set filtrations of a distance function. At each level ε, the ε-sublevel set of the distance d, d_ε, is defined as

$$d_\varepsilon = \bigcup_{i=1}^{n} B(X_i, \varepsilon),$$

where $B(X_i, \varepsilon) = \{x \mid d_{\mathbb{X}}(x, X_i) \leq \varepsilon\}$ denotes a ball of radius ε and center X_i, and $d_{\mathbb{X}}$ is an arbitrary distance function. The metric $d_{\mathbb{X}}$ can be used to enforce some desired property, for example Chazal, Fasy, et al. (2014a) define a distance function, the *Distance to Measure* to *robustify* the estimate. d_ε, usually called the *cover* of $\mathbb{X}_n$ is an approximation of the unknown $\mathbb{X}$ that retains more topological information than the original point–cloud (Figure 2.1).

The topology of d_ε coulf be investigated by computing its Homology Groups, however it is extremely sensible to the radius ε. For each value of ε, in fact, we obtain a different estimate d_ε, with a different topological structure. As shown in Figure 2.2, when ε is small, d_ε is topologically equivalent to $\mathbb{X}_n$: it consists of many connected components but no loops, voids or other higher dimensional structures. Letting ε grow, balls in the cover start to intersect "giving birth" to more complex features, such as cycles. Gradually, increasing ε causes connected components to merge and loops to be filled so that eventually d_ε is topologically equivalent to a ball (or in other words, *contractible*) and again retains no information.

The key feature of encoding data into a filtration is that as ε grows, different sublevel-sets d_{ε_1}, d_{ε_2} are related, so that if a feature is present in both we can say that it remains alive in the interval $[\varepsilon_1, \varepsilon_2]$. Persistent Homology then allows to see how features appear and disappear at different scales. Values ε_b, ε_d of ε corresponding to when two components are connected for the first time (*birth–step*) and when they are connected to some other larger component (*death–step*) are the generators of a Persistent Homology Group.

In the statistical literature, d_ε is often known as the *Devroye–Wise support estimator* (Devroye and Wise 1980). The consistency of the Devroye-Wise estimator justifies and motivates the use of the distance function: as d_ε is a consistent estimator of $\mathbb{X}$, the topology of d_ε is a reasonable approximation of the topology of $\mathbb{X}$.

Kernel Density estimators The second way of linking levelset filtrations and the topology of the support of the distribution generating the data, $\mathbb{X}$, is that the super–levelsets of a density function p can be topologically equivalent to the support of the distribution itself as shown in B. T. Fasy et al. (2014). More formally, if the data are sampled from a distribution P supported on $\mathbb{X}$, and if the density p of P is smooth and bounded away from 0, then there is an interval $[\eta, \delta]$ such that the super–levelset $p^\varepsilon = \{x \mid p(x) \geq \varepsilon\}$ is homotopic (i.e. topologically equivalent) to $\mathbb{X}$, for $\eta \leq \varepsilon \leq \delta$.

Since the true generating density p is most often unknown, it is typically approximated by a kernel density estimator $\widehat{p}$. A naive way to estimate the topology of $\mathbb{X}$ is hence to compute topological invariants of the superlevel set of the kernel density estimator $\widehat{p}$:

$$\widehat{p}^\varepsilon = \{x \mid \widehat{p}_n(x) \geq \varepsilon\}.$$

The superlevel sets $\widehat{p}^\varepsilon$, with $\varepsilon \in [0, \max \widehat{p}]$, form a decreasing filtration, which means that $\widehat{p}^\varepsilon \subset \widehat{p}^\delta$ for all $\delta \leq \varepsilon$. As in the case of distances, for each element in the filtration, i.e. for each value ε, we obtain a different estimate $\widehat{p}^\varepsilon$, whose topology can be characterized by its Homology Groups. Since in practice it is not possible to determine the interval $[\eta, \delta]$ in which the topology of $\widehat{p}^\varepsilon$, is closest to that of $\mathbb{X}$, we analyse the evolution of the topology in the whole filtration. Once again, Persistent Homology allows to analyze how those Homology Groups change with ε. Persistent loops in $\widehat{p}^\varepsilon$ naturally represent circular structures in $\widehat{p}$, Persistent Homology Groups of dimension 2 indicate holes in $\widehat{p}$ and so on.

Far from being trivial, topological features of dimension 0, or connected components have also a relevant interpretation in terms of "bumps". As can be seen from Figure 2.4, connected components in the filtration $\widehat{p}^\varepsilon$, are in fact local maxima of $\widehat{p}$; this is true for any super–levelset filtration. When the filtration is defined in terms of sub–levelset instead, as in the case of the distance function, connected components represent local minima instead.

Regardless of the class of functions f chosen to build the Persistent Homology Groups, the intuition behind the Persistence approach is that features of the real object of interest $\mathbb{X}$ should be found at many different resolutions of its approximation f^{ε}; if a feature of f^{ε} has a "long life", is likely to be a relevant feature of $\mathbb{X}$ as well.

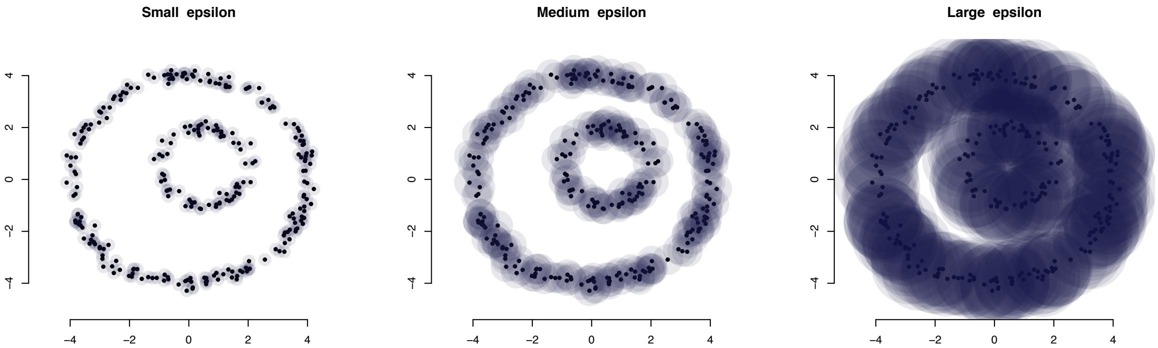

Figure 2.2: From left to right: the cover d_{ε} for the data shown in Figure 2.1 for increasing values of ε. When the value of ε is very small (left) d_{ε} does not have either of the two loops of $\mathbb{X}$. For larger ε (right), the smaller loop in the middle is filled by and the cover d_{ε} only retain the larger loop of $\mathbb{X}$.

2.1.2 Persistent Homology Groups - Formally

The notion of "lifetime" of topological features is formalized as Persistent Homology Groups, a *multiscale* version of Homology Groups that analyses the evolution of the topology of the elements of a filtration. Intuitively, a Persistent Homology Group $H_{k,\delta-\eta}(f)$ of dimension k, consists of the k-dimensional homology classes of f_{η} which are still alive at f_{δ}[1].

The main feature of the cover is that $\forall \delta \leq \eta$, f_{δ}, and f_{η}, are related by inclusion: $f_{\delta} \subseteq f_{\eta}$, which allows to track the evolution of each feature and to see when it appears and disappears.

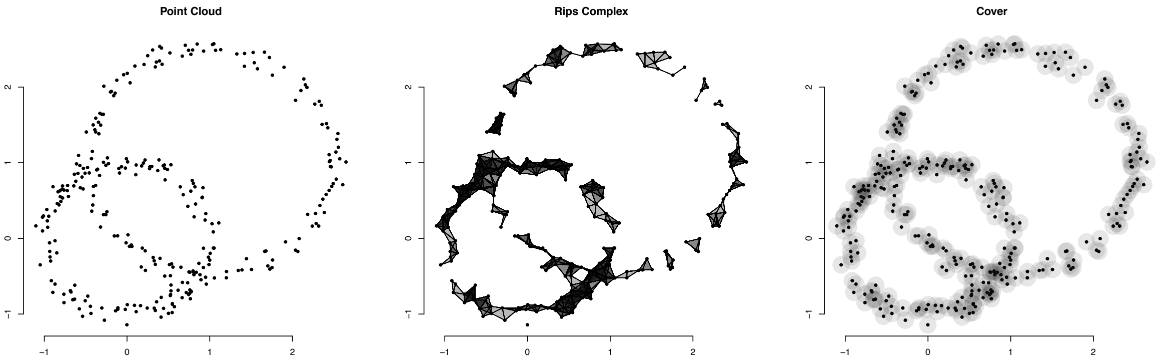

Figure 2.3: From left to right: Data $\mathbb{X}$, Rips complex $Rips_{\varepsilon}(\mathbb{X})$ and corresponding cover $\mathbb{X}^{\varepsilon}$.

At every level ε of the filtration $\mathcal{F}$, Homology Groups of dimension k identify topological features of dimension k; in order to understand which k–dimensional feature survives between

[1] in the following we will refer to sublevel set filtrations and covers but results hold for any kind of filtration.

η and δ, it is necessary to build the map

$$h : H_k(f^{\eta}) \mapsto H_k(f^{\delta}),$$

that shows how Homology Groups at η and δ are related. However, since Homology is a *functor*, it induces a linear map $H(i_{\eta}^{\delta}) : H(f_{\eta}) \mapsto H(f_{\delta})$ on the inclusion map of the $f_{\eta} \hookrightarrow f_{\delta}$, so that $h = H(i_t^s)$.

Definition 2.1 (Persistent Homology Groups)**.** Given a filtration $\mathcal{F} = \{\mathbb{X}_n^{\varepsilon}\}_{\varepsilon}$ indexed on $\mathbb{R}$, i.e. a sequence of topological spaces f_{ε} for each $\varepsilon \in \mathbb{R}$ and maps $f_{\eta} \hookrightarrow f_{\delta}$ for $\eta \leq \delta$, there are natural maps

$$H(i_{\eta}^{\delta}) : H_k(f_{\eta}) \mapsto H_k(f_{\delta}),$$

induced by functoriality. The dimension–k Persistent Homology Group $H_{k,p}$, where $p = \delta - \eta$, are defined as the image of the induced map $H(i_{\eta}^{\delta})$.

From a computational point of view, Persistent Homology Groups can be computed by (means of) simple matrix reduction algorithms, due to the fact that the cover f_{ε} can be approximated by a family of simplicial complexes without loosing any topological information. The most intuitive discrete approximation of the cover f_{ε} is its *Nerve*, also known as Cech complex.

Definition 2.2 (Cech Complex)**.** Given a metric space $(\mathbb{X}, d_{\mathbb{X}})$ the Cech complex $Cech_{\alpha}(\mathbb{X})$ is the set of simplices $\sigma = [X_1, \dots, X_k]$ such that the k closed balls $B(X_i, \alpha)$ have a non empty intersection.

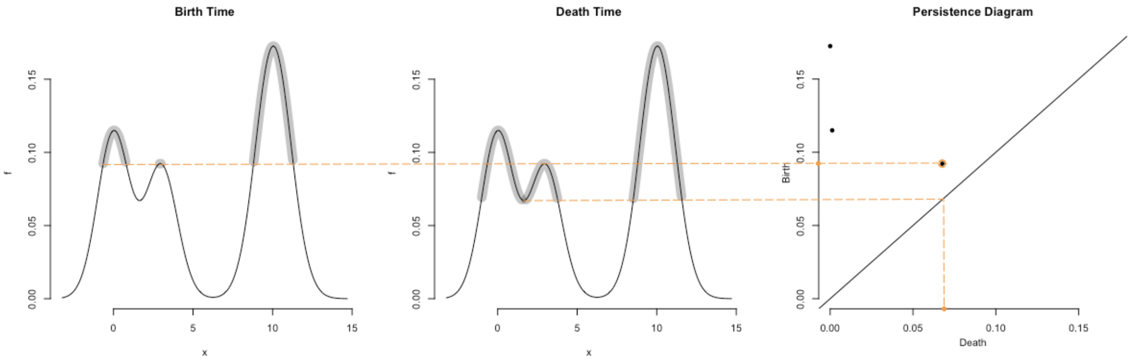

Figure 2.4: From left to right: birth of the smallest peak in the filtration, $\widehat{p}_n^b$, death of the circle $\widehat{p}_n^d$ and summarizing Persistence Diagram.

Since the elements of $\mathbb{X}_n^{\varepsilon}$ are by definition contractible, $\mathbb{X}_n^{\varepsilon}$ is what is called a *good cover* and it satisfies the assumption of the *Nerve Theorem*.

Theorem 2.1 (Nerve)**.** *A good cover and its nerve are homotopic.*

The Nerve Theorem implies that the homology group of $Cech_{\varepsilon}$ are topologically equivalent to those of f^{ε}. Nevertheless, computing the Cech complex itself can still be computationally challenging; for this reason the Vietoris–Rips complex, another combinatorial representation of f^{ε}, is typically preferred.

Definition 2.3 (Vietoris–Rips Complex)**.** Given a metric space $(\mathbb{X}, d_{\mathbb{X}})$ the Vietoris–Rips complex $Rips_\alpha(\mathbb{X})$ is the set of simplices $\sigma = [X_1, \ldots, X_k]$ such that $d_{\mathbb{X}}(X_i, X_j) \leq \alpha$ for all i, j.

Even though the Nerve theorem does not hold for Vietoris–Rips complexes, its topology is still close to the one of f^ε due to its proximity to the Cech complex:

$$Rips_\varepsilon(\mathbb{X}) \subseteq Cech_\varepsilon(\mathbb{X}) \subseteq Rips_{2\varepsilon}(\mathbb{X}).$$

Other families of simplicial complexes such as Delauney Triangulations, Witness complex or Alpha shapes can be used to compute Persistent Homology as well. A formal model selection procedure for selecting the best family of simplicial complexes in this context, however, is non trivial, as, besides computational tractability, it may not be obvious to formalize desirable properties of a simplicial complex. There are ways of choosing the best simplicial complex representation of a space, however, for example Caillerie and Michel (2011) builds a penalized Risk, using metric entropy as a penalty. The risk is defined from L_2, so that the result is a penalized Least Square Error.

2.2 Persistence Diagrams

The evolution of the topology of f_ε can be summarized by the *Persistence Diagram* D, a multiset whose generic element $x_i = (b_i, d_i)$ is the $i^{\texttt{th}}$ feature in f_ε (or equivalently, the $i^{\texttt{th}}$ generator of a Persistent Homology Group). The first coordinate, the "birth time" b_i, represent how soon in the filtration the $i^{\texttt{th}}$ feature appears, i.e. the first value ε for which the $i^{\texttt{th}}$ feature can be found in f_ε; the second coordinate, the "death time"" d_i, represents when the feature disappear, i.e. the first value ε for which f_ε does not retain the $i^{\texttt{th}}$ feature anymore. Since two or more feature can share birth and death time, each point has multiplicity equal to the number of features, except for the diagonal, whose points have infinite multiplicity. As death always occurs after birth, all points in the diagram are in or above the diagonal. The *Persistence Barcode* is an equivalent representation, where each bar is a feature whose length correspond to the lifetime of the corresponding feature.

The "lifetime" or *persistence* $\text{pers}(x) = d - b$, of a feature can be considered as a measure of its importance. Points that are close to the diagonal represent features that appear and disappear almost immediately and may be neglected; a diagram whose only elements are the points of the diagonal $D_\emptyset$ is said to be *empty*. Figure 2.5 shows the Persistence Diagrams corresponding to the point–cloud shown in Figure 2.1. If we focus on the red elements, which represent the 1-dimensional features (or loops), we can see how the structure of the unknown $\mathbb{X}$ is captured by the diagram. Not only the two loops of $\mathbb{X}$ are clearly recognizable both from the Persistence Diagram (the two triangles above the diagonal) and the Persistence Barcode (the two longer lines), but it also possible to distinguish between the two of them, as the the small one is in fact slightly less persistent than the larger one.

Although in theory a feature may never die, in diagrams built from point clouds all the information is contained between the diagonal and the diameter of the data. For the sake of simplicity, we thus limit our analysis to bounded diagrams, i.e. diagrams without infinitely persistent features.

Definition 2.4 (Space of Persistence Diagrams)**.** Let $\mathrm{Pers}_p(D) = \sum_{x \in D} \mathrm{pers}(x)^p$ be the degree–p *total persistence* of a persistence diagram d. Define the space of persistence diagrams $\mathcal{D}$ as

$$\mathcal{D} = \{D \mid \mathrm{Pers}_p(D) < \infty\},$$

where $D_\emptyset$ is the persistence diagram containing only the diagonal.

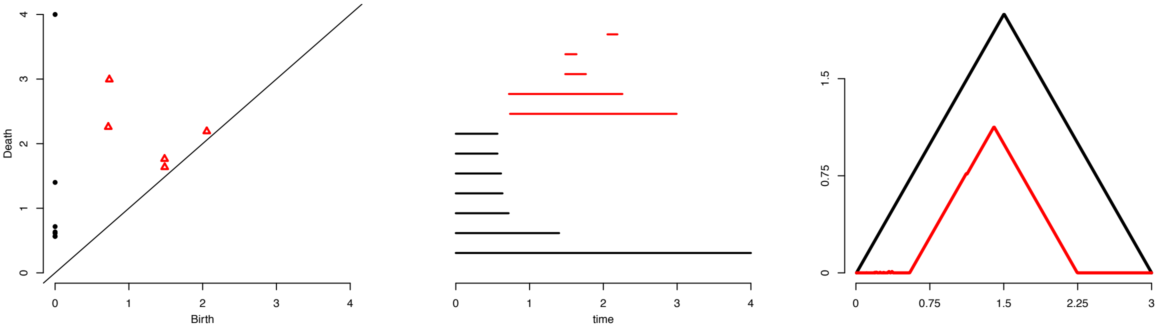

Figure 2.5: (From left to right) Persistence Diagram, Persistence Barcode and Persistence Landscape of data shown in Figure 2.1. In black 0–dimensional features (connected components), in red 1–dimensional features (loops).

As Homology, and hence Persistent Homology Groups, can be defined for every dimension k, so does the Persistence Diagram. However, it is worth stressing that diagrams corresponding to different dimensions are shown together for visualization purposes only and must be considered separately in inferential procedures.

2.2.1 Metrics for Persistence Diagrams

Persistence diagrams can be compared through several metrics, most noticeably the Bottleneck and the Wasserstein distance, which add to $\mathcal{D}$ the structure of a metric space. The Wasserstein distance, also known as Earth Mover distance or Kantorovich distance is a popular metric in Probability and Computer Science as well as Statistics.

Definition 2.5 (Wasserstein Distance between Persistence Diagrams)**.** Given a metric d, called *ground distance*, the Wasserstein distance between two persistence diagrams D and D' is defined as

$$W_{d,p}(D, D') = \left[\inf_{\gamma} \sum_{x \in D} d(x, \gamma(x))^p \right]^{\frac{1}{p}},$$

where the infimum is taken over all bijections $\gamma : D \mapsto D'$.

Depending on the choice of the ground distance d, Definition 2.5 defines a family of metrics, whose most prominent member in `TDA` literature is the L^∞–Wasserstein distance, W_{L^∞}, defined as:

$$W_{L^\infty,p}(D, D') = \left[\inf_{\gamma} \sum_{x \in D} \|x - \gamma(x)\|_\infty^p \right]^{\frac{1}{p}}.$$

When $p = \infty$, the distance $W_{L^\infty,\infty}$ defined as

$$W_{L^\infty,\infty}(D, D') = \inf_{\gamma} \sup_{x \in D} \|x - \gamma(x)\|_\infty,$$

takes the name of *Bottleneck distance.*

Despite being less popular in the TDA framework, another important choice of ground distance is the L^2–norm, especially in the case $p = 2$, for which Turner et al. (2014) proved that $W_{L^2,2}$ is a geodesic on the space of persistence diagrams.

Proposition 2.1 (Turner et al.)**.** The space of Persistence Diagrams $\mathcal{D}$ endowed with $W_{L^2,2}$ is a geodesic space.

The space $\mathcal{D}$ is separable and complete in both W_{L^∞} and W_{L^2}, hence is a Polish Space Mileyko, Mukherjee, and Harer (2011).

2.2.2 Stability

Defining metrics on $\mathcal{D}$ allows for a notion of stability (Chazal et al. 2012), which, roughly speaking, states that similar topological spaces must have similar diagrams.

Theorem 2.2 (Chazal et al.)**.** *Let f and g be two functions on a triangulable space $\mathbb{X}$ and let D_f, D_g be the Persistence Diagram built on their respective sublevel (or superlevel) set filtrations, then*

$$W_{L^\infty,\infty} \leq \|f - g\|_\infty,$$

where $\|f\|_\infty = \sup_x |f(x)|$ is the L^∞–norm.

In the special case of $f = d_{\mathbb{X}}$ and $g = d_{\mathbb{Y}}$ two distance functions defined on two point–clouds $\mathbb{X}$ and $\mathbb{Y}$ respectively, the stability result can be written in a more easily interpretable way:

$$d_B\left(D_{\mathbb{X}}, D_{\mathbb{Y}}\right) \leq 2\, d_{GH}\left(\mathbb{X}, \mathbb{Y}\right),$$

where $d_G H(\mathbb{X}, \mathbb{Y})$ is the *Gromov–Hausdorff* distance between two topological spaces $\mathbb{X}$ and $\mathbb{Y}$. Stability is a core result in TDA for two reasons:

- *the persistence diagram is a topological signature:* stability reassures us that if two point-clouds $\mathbb{X}, \mathbb{Y}$ are similar their Persistence Diagrams will be as well, and is therefore instrumental for using them in statistical tasks such as classification or clustering;

- *the persistence diagram is statistically consistent:* stability reassure us that if we are using a point–cloud $\mathbb{X}_n$ to estimate the topology of an unknown object $\mathbb{X}$, if $\mathbb{X}_n \to \mathbb{X}$ as $n \to \infty$, then $D_{\mathbb{X}_n}$ converges to $D_{\mathbb{X}}$ as well.

Stability is also a key result for distinguishing topological noise from topological signal. Building on the core idea that features that are close to the diagonal are more likely to be noise than those that are far away from it, B. T. Fasy et al. (2014) proposes a way to define a bootstrap confidence band around the diagonal. Points of the diagram laying outside the band are most likely signal, whereas those inside the band may be just noise. It is worth noticing however, that as such bands are defined using the right-hand-side term of the inequality, they are typically rather conservative.

2.3 Persistence Landscape

Persistence Diagrams are defined in spaces endowed with only a metric structure, which can be limiting in data analysis. A collection of Persistence Diagrams $D_1, \dots, D_n$ in fact does not have a unique mean, nor a satisfying measure of variability Turner et al. (2014). More critically, although it is possible to define a probability distribution on the space $\mathcal{D}$ (Mileyko, Mukherjee, and Harer 2011), it is still not clear how to explicitly derive it (if it is possible to derived it at all). In order to overcome these issues and to work with more statistics-friendly spaces, several tools have been developed to convert Persistence Diagrams into functional objects, the most famous being the Persistence Landscape (Bubenik 2015) and the Persistence Silhouette (Chazal, Fasy, et al. 2014b). These topological summaries are built by mapping each point $x = (b, d)$ of a Persistence Diagram D to a piecewise linear function called the "triangle" function T_x, which is defined as:

$$T_x(t) = \begin{cases} t - b + d & t \in [b-d, b], \\ b + d - t & t \in (b, b+d], \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

Informally a triangle function links each point of the diagram to the diagonal with segments parallel to the axes, and then rotates them of 45 degrees.

The triangles T_x can be combined in many different ways. If we take their k-max, i.e. the $k^{\texttt{th}}$ largest value in the set $T_x(y)$, we obtain the $k^{\texttt{th}}$ *Persistence Landscape*

$$\lambda_D^k(y) = k\text{-}\max_{z \in D} T_x(y) \qquad k \in \mathbb{N}^+.$$

The Persistence Landscape λ_D is a representation of the Persistence Diagram D as a collection $\{\lambda_D^1, \dots, \lambda_D^K\}$ of piecewise linear functions, indexed by the order of the maximum to be considered in defining the landscape, k. If we take the weighted average of the functions $T_z(y)$, we have the *Power Weighted Silhouette*

$$\psi_p(t) = \frac{\sum_{x \in D} w_x^p T_x(y)}{\sum_{x \in D} w_x^p}.$$

While the space of Persistence Diagrams $\mathcal{D}$ is only a metric space, Persistence Landscapes are defined in a much richer Banach space $\mathcal{L}$, endowed with the following norm

$$\|\lambda_D\|_p^p = \sum_k \left\|\lambda^k\right\|_p^p,$$

where $\left\|\lambda^k\right\|_p$ is the L^p–norm

$$\left\|\lambda^k\right\|_p = \left(\int \lambda^k \mathrm{d}\mu\right)^{1/p}.$$

It is not possible to go back from Persistence Landscapes to Persistence Diagrams, meaning that there is a loss of information in going from Persistence Diagrams to Persistence Landscapes. However the Persistence Landscape is still informative, since stability continues to hold.

Theorem 2.3 (Bubenik)**.** *Let f, g be two functions on $\mathbb{X}$ and let D_f and D_g be the Persistence Diagrams built from their superlevel (or sublevel) sets, then*

$$d_\Lambda\left(\lambda_{D_f}, \lambda_{D_g}\right) \leq \|f - g\|_\infty ,$$

where $d_\Lambda\left(\lambda_{D_f}, \lambda_{D_g}\right) = \left\|\lambda_{D_f} - \lambda_{D_g}\right\|_\infty$ is the L^∞–distance in the space of Persistence Landscapes, $\mathcal{L}$.

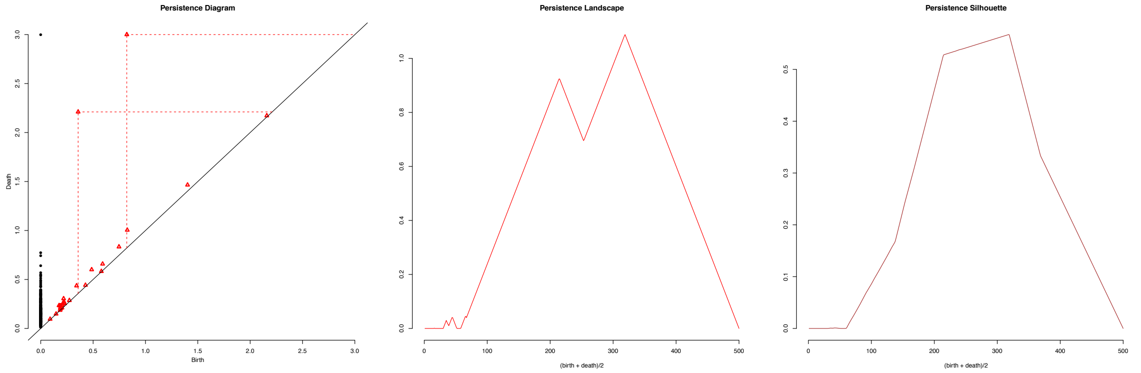

Figure 2.6: A Persistence Diagram (left) and its corresponding Persistence Landscape (center) and Persistence Silhouette (right).

Persistence Landscapes are piece–wise linear functions, which makes it possible to define a (unique) mean and a variance for any collection of them. The main advantage of the Persistence Landscape over the Persistence Diagram is that it is defined in a Banach Space, which is instrumental in statistical learning as it allows for a full characterization of the Persistence Landscape as a random variable

2.3.1 Probability in Banach Spaces / A modicum

In order to better understand the desirable properties of topological summaries defined in a Banach space rather than in just a metric one, we quickly review the basic of Probability in Banach spaces; a more complete overview can be found in Ledoux and Talagrand (2013). Let $\mathcal{B}$ be a real, separable Banach space with norm $\|\cdot\|$. Let $(\Omega, \mathcal{F}, \mathcal{B})$ be a probability space and let

$$V : (\Omega, \mathcal{F}, \mathcal{B}) \mapsto \mathcal{B},$$

be a Borel random variable with values in $\mathcal{B}$.

We call an element of $\mathcal{B}$ the *Pettis integral* of V if $\mathbb{E}(f(V)) = f(\mathbb{E}(V))$ for all $f \in \mathcal{B}^\star$, where $\mathcal{B}^\star$ is the space of continuous linear real–valued functions on $\mathcal{B}$, i.e. the topological dual space of $\mathcal{B}$. The *Pettis integral* is the analogous of the expected value for a $\mathcal{B}$–valued random variable. The following proposition gives us a sufficient condition for its existence.

Proposition 2.2. If $\mathbb{E}\,\|V\| < \infty$, then V has a *Pettis integral* and $\|\mathbb{E}(V)\| \leq \mathbb{E}\,\|V\|$.

Notice that $\|V\|$ is a real valued random variable.

The Pettis integral can be used to define an extension of the Law of Large numbers for a $\mathcal{B}$–valued random variable. Recall that for a sequence $\{Y_n\}_n$ of $\mathcal{B}$–valued random variables:

- $\{Y_n\}_n$ converges *almost surely* to a $\mathcal{B}$–valued random variable Y if $\mathbb{P}(\lim_{n\to\infty} Y_n) = 1$.
- $\{Y_n\}_n$ converges *weakly* to a $\mathcal{B}$–valued random variable Y if $\lim_{n\to\infty} \mathbb{E}(\phi(Y_n)) = \mathbb{E}(\phi(Y))$ for all bounded continuous functions $\phi : \mathcal{B} \mapsto \mathbb{R}$.

Theorem 2.4 (Strong Law of Large Numbers)**.** *Let $\{V_n\}_{n\in\mathbb{N}}$ be a sequence of independent copies of V and, for a given n, let $S_n = V_1 + \cdots + V_n$,*

$$\frac{S_n}{n} \to \mathbb{E}(V) \quad \textit{almost surely} \iff \mathbb{E}\,\|V\| < \infty.$$

There is an extension of the Central Limit Theory as well, which states the convergence to a Gaussian random variable. In a Banach Space $\mathcal{B}$, a random variable G is said to be *Gaussian* if for each $f \in \mathcal{B}^\star$, $f(G)$ is a real valued Gaussian random variable with 0 mean. The covariance structure of a $\mathcal{B}$–valued random variable, which fully characterize a Gaussian Random Variable in a Banach Space, is given by

$$\mathbb{E}\big[(f(V) - \mathbb{E}[f(V)]) \cdot (g(V) - \mathbb{E}[g(V)])\big],$$

where $f, g \in \mathcal{B}^\star$.

Theorem 2.5 (Central Limit Theorem)**.** *Assume $\mathcal{B}$ has type 2. If $\mathbb{E}(V) = 0$ and $\mathbb{E}(\|V\|^2) < \infty$ then $\frac{S_n}{\sqrt{n}}$ converges weakly to a Gaussian random variable $G(V)$ with the same covariance structure as V.*

The extension of these two result to the case of Persistence Landscapes is immediate.

2.4 Persistence Flamelets

Persistence Diagrams and Persistence Landscapes gives us a full characterization of any function of data f in terms of the topology of its sub–levelset (or super–levelset) filtration, however they do allow f to vary. In this section we focus on the case where rather than one function f we are dealing with a family of functions $\mathcal{F} = \{f(\cdot;\sigma),\ \sigma \in [0,1]\}$, indexed by some parameter σ[2], which represent the resolution or the scale of the object $f(\cdot;\sigma)$. This is a challenging yet common framework in statistics, where scale dependent tools are already almost ubiquitous (smoothers being the most trivial example of it). Due to the ever–growing complexity of data, being able to examine it at different resolutions, hence obtaining different insights, has in fact become a crucial feature of statistical tools, however summarizing the information coming from different scales is non trivial.

Although traditional methods focus on selecting the optimal scale σ^*, inspired by scale space theory, we adopt the idea that there is no "real" resolution, but, as different scales yield different information, all of them must be simultaneously take into account. We restrict ourselves to the case where the $\mathcal{F} = \{f(\cdot;\sigma),\ \sigma \in [0,1]\}$ is continuously indexed by the scale parameter σ. Example of this that we will explore more thoroughly in the following are kernel smoothers, for which the resolution σ is given by the bandwidth parameter h, and time–varying processes, whose scale σ is time, t.

[2] For the sake of simplicity we will assume $\sigma \in [0,1]$, but we only require σ to be bounded.

Previous attempts at encoding a multi-resolution family $\mathcal{F} = \{f(\cdot;\sigma),\ \sigma \in [0,1]\}$ into the TDA framework is to consider the Persistence Diagram itself as a function of the scale parameter σ. The family of Persistence Diagrams $\mathbb{D} = \{D_\sigma,\ \sigma \in [0,1]\}$ corresponding to $\mathcal{F}$, is known as *Persistence Vineyards* (Cohen-Steiner, Edelsbrunner, and Morozov 2006) and is a stable and continuous representation of the topology of the whole $\mathcal{F}$, as shown in Morozov (2008). Persistence Vineyards, however, share all the drawbacks and limitations of Persistence Diagrams, more specifically they lack a unique average and a measure of variability for a group of them (Turner et al. 2014), and, once again, since it is not yet clear whether or not it is possible to explicitly define a probability distribution on the space of Persistence Vineyards, their use in statistical inference is severely compromised (Mileyko, Mukherjee, and Harer 2011).

Building on TDA's toolbox, and Persistence Landscapes in particular, we introduce a new topological summary, the Persistence Flamelets, which overvcomes most of these issues while still being able to characterize both the topology at each resolution $f(\cdot;\sigma)$ and how it changes with σ. The Persistence Flamelets is an easily interpretable tool, it allows for visualization of arbitrarily high dimensional features and is a stable topological signature.

It is worth noticing that although in the following we focus on Persistence Landscapes, the same results hold for Silhouettes as well. In order to explicitly take into account the multiple resolutions of $\mathcal{F}$, we consider the Persistence Landscapes λ_{D_σ} corresponding to the family $\mathcal{F} = \{f(\cdot;\sigma),\ \sigma \in [0,1]\}$ as a function of the scale parameter σ. Visually we can think of such function as a "flow" of landscapes, one for each resolution, smoothly moving and resembling a tiny fire (see, for example, Figure 2.10).

Definition 2.6 (Persistence Flamelets)**.** Given a collection of Persistence Diagrams D_σ, continuously indexed by some parameter $\sigma \in [0,1]$, and $k \in \mathbb{N}^+$, we define the $k^{\mathtt{th}}$ *Persistence Flamelets* as the function

$$\Lambda^k(\sigma, y) = \lambda^k_{D_\sigma}(y) \qquad \forall\, \sigma \in [0,1],\ y \in \mathbb{R},\ k \in \mathbb{N}^+.$$

As the Landscape itself, the *Persistence Flamelets* Λ is also a collection $\Lambda = \{\Lambda^{(k)},\ k \in \mathbb{N}^+\}$ indexed by the order of the max we consider. The theoretical reassurance that the Persistence Flamelets is a meaningful topological summary is its stability, which we will prove in the following. Before doing so, however, we need to introduce a notion of *proximity* between Persistence Flamelets.

Definition 2.7 (Integrated Landscape distance)**.** Let $\mathbb{D} = \{D_\sigma,\ \sigma \in [0,1]\}$, $\mathbb{E} = \{E_\sigma,\ \sigma \in [0,1]\}$ two Persistence Vineyards and $\Lambda_\mathbb{D}, \Lambda_\mathbb{E}$ the corresponding Persistence Flamelets. We define the *Integrated Landscape distance* between $\Lambda_\mathbb{D}$ and $\Lambda_\mathbb{E}$ as

$$I_\Lambda(\Lambda_\mathbb{D}, \Lambda_\mathbb{E}) = \int_0^1 d_\Lambda(\lambda_{D_\sigma}, \lambda_{E_\sigma})\,\mathrm{d}\sigma.$$

Theorem 2.6. *Let $\mathbb{D} = \{D_\sigma,\ \sigma \in [0,1]\}$, $\mathbb{E} = \{E_\sigma,\ \sigma \in [0,1]\}$ two Persistence Vineyards and $\Lambda_\mathbb{D}, \Lambda_\mathbb{E}$ the corresponding Persistence Flamelets, then:*

1. *$\Lambda_\mathbb{D}$ and $\Lambda_\mathbb{E}$ are continuous with respect to the Bottleneck distance;*

2. $I_\Lambda(\Lambda_\mathbb{D}, \Lambda_\mathbb{E}) \leq I_B(\mathbb{D}, \mathbb{E})$

where $I_B(\mathbb{D}, \mathbb{E}) = \int_0^1 d_B(D_\sigma, E_\sigma)\, dt$ *is the Integrated Bottleneck distance for Persistence Vineyards as defined in @Munch2013.*

The proof is a direct consequence of the Stability Theorem for Persistence Landscapes (Theorem 2.3) and the continuity of Persistence Vineyards, in fact:

1. For a fixed σ, consider D_σ and $D_{\sigma+\varepsilon}$ (same applies for $\mathbb{E}$). By 2.3 and the continuity of $\mathbb{D}$ we have
$$0 \leq \lim_{\varepsilon \to 0} d_\Lambda \left(\lambda_{D_\sigma}, \lambda_{D_{\sigma+\varepsilon}}\right) \leq \lim_{\varepsilon \to 0} d_B \left(D_\sigma, D_{\sigma+\varepsilon}\right) = 0.$$

2. Since for a fixed σ we have, by Theorem 2.3 we have
$$d_\Lambda \left(\lambda_{D_\sigma}, \lambda_{E_\sigma}\right) \leq d_B \left(D_\sigma, E_\sigma\right)$$
integrating both terms is enough to prove the result.

The Persistence Flamelets is also a random variable defined in a Banach space. In analogy with what Bubenik (2015) has done for Persistence Landscapes, we define a norm for Persistence Flamelets, more specifically

$$\|\Lambda\|_p^p = \int_0^1 \sum_k \left\|\lambda^k(t)\right\|_p^p \mathrm{d}t$$

Then following Ledoux and Talagrand (2013), we can extend the Law of Large Numbers and the Central Limit Theorem to this new object.

Corollary 2.1 (Strong Law of Large Numbers)**.** Let $\{\Lambda_n\}_{n\in\mathbb{N}}$ be a sequence of independent copies of Λ and, for a given n, let $S_n = \Lambda_1 + \cdots + \Lambda_n$, where the sum is defined pointwise.

$$\frac{S_n}{n} \to \mathbb{E}(\Lambda) \quad \text{almost surely} \iff \mathbb{E}\,\|\Lambda\| < \infty.$$

Corollary 2.2 (Central Limit Theorem)**.** Assume $\mathcal{B}$ has type 2. If $\mathbb{E}(V) = 0$ and $\mathbb{E}(\|\Lambda\|^2) < \infty$ then $\frac{S_n}{\sqrt{n}}$ converges weakly to a Gaussian random variable $G(\Lambda)$ with the same covariance structure as Λ.

Proofs directly follow from Theorem 2.4 and Theorem 2.5.

2.4.1 Some intuition / EEG Dynamic Point–Clouds

A short example will clarify when this object, until now rather abstract, may be encountered and fruitfully used. The easiest way to understand the need for topological characterization of a continuously varying space is to consider the case where the scale parameter σ is time, t. The Persistence Flamelets allows in fact for a characterization of a time–varying system $\mathcal{F} = \{f(\cdot; t),\ t \in [0, 1]\}$ in terms of its topology (Munch et al. 2015, Munch (2013)) by allowing us to simultaneously study the shape of any time–dependent function f_t and how it evolves with time t.

Again, although this framework is general enough to cover any arbitrary function $f(\cdot; t)$, as long as it is continuous with respect to time, we are especially interested in the case where $f(\cdot; t)$ is a function of data.

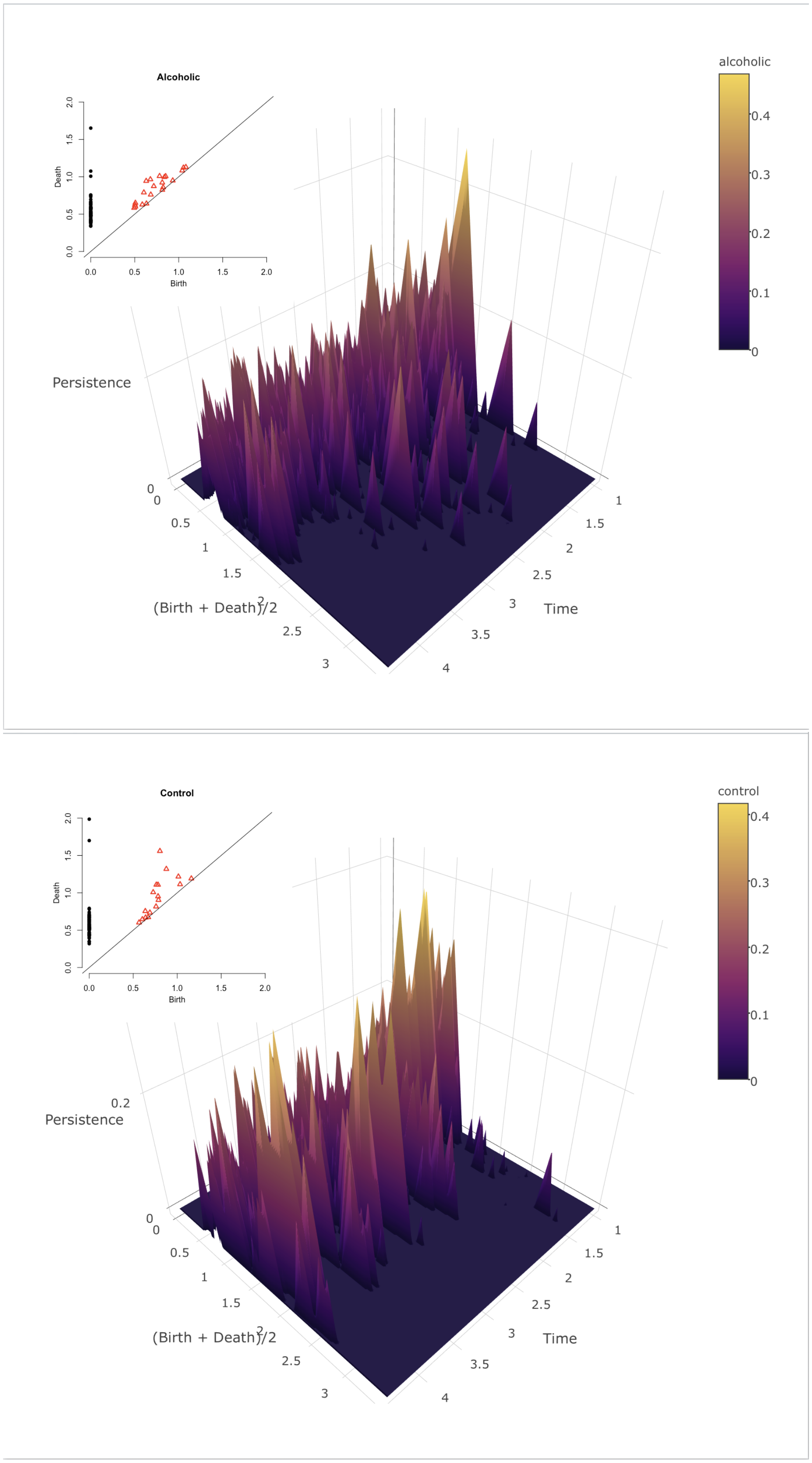

Figure 2.7: Persistence Flamelets of Dimension 1 for the EEG data of one alcoholic (upper) and one control (lower) subject.

Assume that at each time t we observe a sample $\mathbb{X}(t) = \{X_1(t), \dots, X_k(t)\}$ drawn from some distribution P_t. The Persistence Flamelets Λ built on distance functions or kernel density estimators estimate the topology of the whole continuous–time generating process $\{P_t,\ t \in [0,1]\}$. The trace of the sample in the time interval $\{\mathbb{X}(t),\ t \in [0,1]\}$, usually called *Dynamic Point Cloud*, is just a high dimensional time series, hence the Persistence Flamelets can be exploited as a tool to extract a new type of insights on time series of arbitrarily high dimension. In the special case of dynamic point–clouds, the stability result of Theorem 2.6 can be restated as follows.

Corollary 2.3. Let $\{\mathbb{X}(t), \mathbb{Y}(t)\}$ with $t \in (0,1)$ two continuous dynamic point clouds, $\Lambda_{\mathbb{X}}$ and $\Lambda_{\mathbb{Y}}$ their corresponding Persistence Flamelets, then:

$$I_{\Lambda}(\Lambda_{\mathbb{X}}, \Lambda_{\mathbb{Y}}) \leq I_H(\mathbb{X}, \mathbb{Y}),$$

where $I_H(\mathbb{X}, \mathbb{Y}) = \int_0^1 d_H(\mathbb{X}(t), \mathbb{Y}(t))\mathrm{d}t$ is the Integrated Hausdorff distance for dynamic point–clouds, as defined in Munch (2013).

Figure 2.7 shows two Persistence Flamelets built from electroencephalography (EEG) tracks, freely available on the `UCI Machine Learning Repository`. EEG are electric impulses recorded at a very high frequency (256 *Hz*) through multiple electrodes (64 in this study), located in different areas of the skull. The topology of EEG data has successfully being investigated by Wang, Ombao, and Chung (2018), who characterized epilepsy in terms of local peaks, through the Persistence Landscape. Their findings are encouraging, as they detect significative difference between the signal during an epilepsy attack and in a control state, however they are limited by the fact that they can consider only one EEG channel at the time, being unable to deal with the spatial and the temporal resolution simultaneously. Both the domains retain relevant information: at each time t, connected components and loops represent area of the brain that share the same behavior, which is relevant information per se, but since it is also important to assess whether or not these connection persist in time, this kind of data fits perfectly in our framework. The Persistence Flamelets highlights differences in the brain's behavior of the two groups, as illustrated by Figure 2.7, which represents the Persistence Flamelets, for one alcoholic and one control patient. The signal from the control patient, in fact, is strongly characterized by a few persistent features. In the alcoholic patient instead there is less structure; there seems to be more features than in the control patient, but they all have a smaller persistence, and could therefore be interpreted as noise.

In order to understand whether this difference is just circumstantial or it may be more grounded, we compare the EEGs of ten alcoholic and ten control patients from the same repository, all subject to the same stimulus. For each of them we have 5 trials of 1 second; EEG are typically very noisy hence we average them across repetition before computing their topological summaries. For this application we compare Flamelets based on the Persistence Silhouette rather than the Persistence Landscape, as the erratic behavior of the loops makes it difficult to choose the order k of the max. We performs a simple permutation test to compare the two groups, using the Integrated Landscape Distance, and results shown that while there may not be significative difference between the dimension 0 Flamelet, the two groups are significatively different when compared through their dimension 1, which could motivate further investigation on the presence and formation of loops in brain activity.

2.5 Data Smoothing / Applications & Comparisons

In the statistical literature, scale–space ideas have been especially popular in the context of *data smoothing.* In its broader definition, data smoothing is a family of methods aimed at recovering some structure in the data. Depending on their scale, however, smoothing methods may enhance noise or neglect relevant features, so that it is crucial to understand the impact of the smoothing level on the smoothed object.

Persistence Flamelets can be used to summarize and evaluate the evolution of the whole smoothing process. The two main features of the Persistence Flamelets is that they allow for an intuitive visualization of the dynamics of the smoothing process and that they can be exploited to track the appearance and disappearance of feature of arbitrary dimension.

Among all the smoothing methods, we focus on Kernel Density Estimation (KDE) (Scott 2015), for which the role of topological features (especially that of $0^{\mathtt{th}}$ dimensional Homology Groups) is a well established problem, see for example Chaudhuri and Marron (1999). Features affected by the smoothing process such as local peaks (or, in topological terms, $0^{\mathtt{th}}$ dimensional Homology Groups), are in fact especially meaningful in the case of KDE; local modes of a density and their basin of attraction represent for example one way of defining clusters (Ester et al. 1996,Comaniciu and Meer (2002)). Persistence Flamelets allows us to explore also higher dimensional features, such as cycles or voids, which have been noticeably neglected.

Given a sample $\{X_1 \ldots, X_n\}$, drawn from some smooth density p, a Kernel Density Estimator $\widehat{p}_h$ is defined as

$$\widehat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i),$$

where $K_h(x-y) = \frac{1}{h}K(\frac{x-y}{h})$ is a scaled kernel, h is the bandwidth parameter and $K(\cdot)$, the kernel, is a non-negative, symmetric function that integrates to 1.

While any kernel function $K(\cdot)$ may be used without compromising the performance of the estimator, the bandwidth parameter represent the level of smoothing and needs to be finely tuned. In the scale-space approach, given some bounded range of bandwidths $H \subset \mathbb{R}^+$, all the estimators $\widehat{p}_h$ are simultaneously considered, so that the object of interest becomes the family of smooths $\mathcal{F} = \{\widehat{p}_h \,:\, h \in H\}$. Since K_h is continuous with respect to h by definition, it is immediate to see that the Persistence Flamelets can be used to investigate and characterize $\mathcal{F}$.

From an exploration perspective, the first attempt at investigating the relation between the bandwidth of a kernel density estimator and its topology was `SiZer` (Chaudhuri and Marron 1999). Roughly speaking, given a sample $\{X_1, \ldots, X_n\}$ drawn from a univariate density p, `SiZer` (SIgnificant ZERo crossings of derivatives) is a map showing where in space, x, and scale, h, the kernel density estimator $\widehat{p}_h(x)$ is significantly increasing or decreasing. Since local peaks of a curve can be thought of as points where its derivative changes sign, the basic idea of `SiZer` is assess where this change happens, by testing whether the sign of the derivative $\widehat{p}'_h(x)$ for each couple of values (x, h) is positive or negative. Values (x, h) corresponding to significantly positive derivatives are shown in blue and significantly negative are shown in red, as in Figure 2.10.

`SiZer` is intrinsically 1–dimensional and even though it has been extended to 2–dimensional densities, especially in the context of image analysis as in Godtliebsen, Marron, and Chaudhuri (2004), the features it hunts for are always and only local modes. The Persistence Flamelets provides a further extension in two different directions:

- it can be used to investigate topological features of any dimension, rather than only feature of dimension 0, i.e. local peaks;
- it does not depend on the dimension of the data and can thus be used to investigate kernel densities for very high dimensional data.

Finally, even though, with respect to `SiZer`, the Persistence Flamelets lacks of statistical testing to asses the significance of each peak, it provides a measure of the relevance of each feature, its persistence.

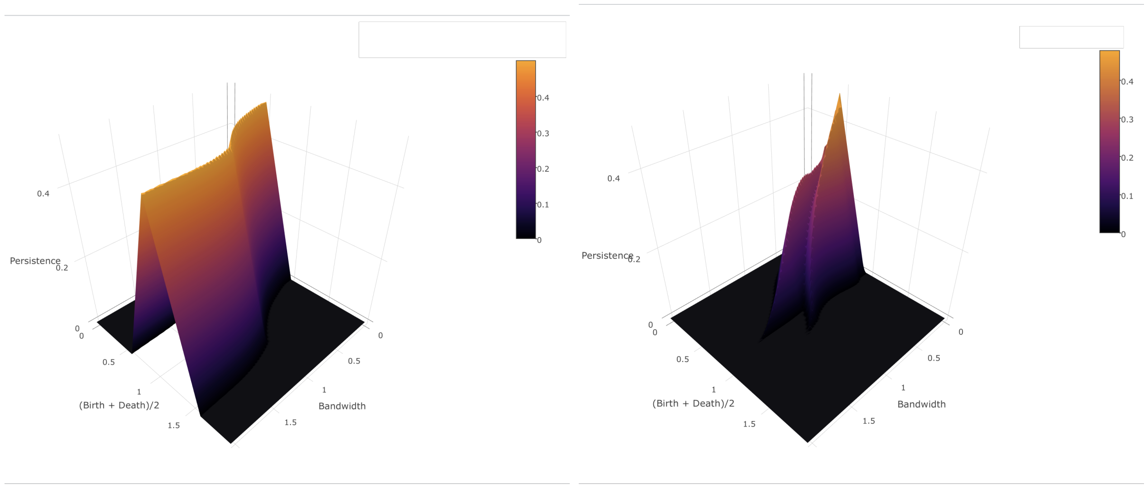

Figure 2.8: $1^{\texttt{st}}$ (left) and $2^{\texttt{nd}}$ (right) Persistence Flamelets of dimension 0.

2.5.1 Bandwidth Exploration

We now show two real–data applications. In the first univariate one we quickly compare the Persistence Flamelets with `SiZer` and show that, when both are available they yield similar insights. The second is a bivariate example, which motivates investigating higher dimensional features and highlights the potential of the Persistence Flamelets when other tools are not available.

Eartquakes I / Depth In our first example we consider a classical dataset in kernel density estimation, the depth of the 512 earthquakes beneath the Mt. St. Helens volcano in the months before the eruption of 1982 (more details can be found in Scott (2015)). Figure 2.8 shows the $1^{\texttt{st}}$ and the $2^{\texttt{nd}}$ Persistence Flamelets for the 0 dimensional topological feature of the density estimator $\widehat{p}$ built with the Gaussian Kernel:

$$\widehat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi h}} \exp\left\{\frac{1}{2h}(x - X_i)^2\right\}.$$

The 1^{st} Persistence Flamelets consists of only one peak, representing the global maximum, which, as we can expect, always persists. This is not very informative, and when analyzing dimension 0 topological features, it is thus advisable to consider 2^{nd} Persistence Flamelets, which represents the most relevant local peaks.

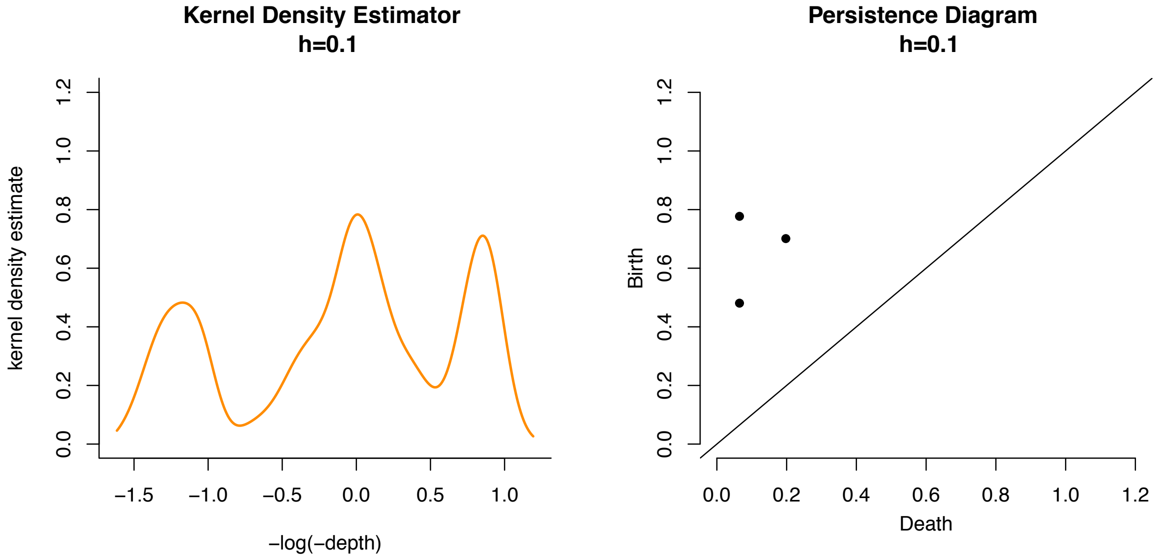

Figure 2.9: From left to right: Kernel Density Estimator of the Mt. St. Helens dept data (with $h = 0.1$) and corresponding Persistence Diagram.

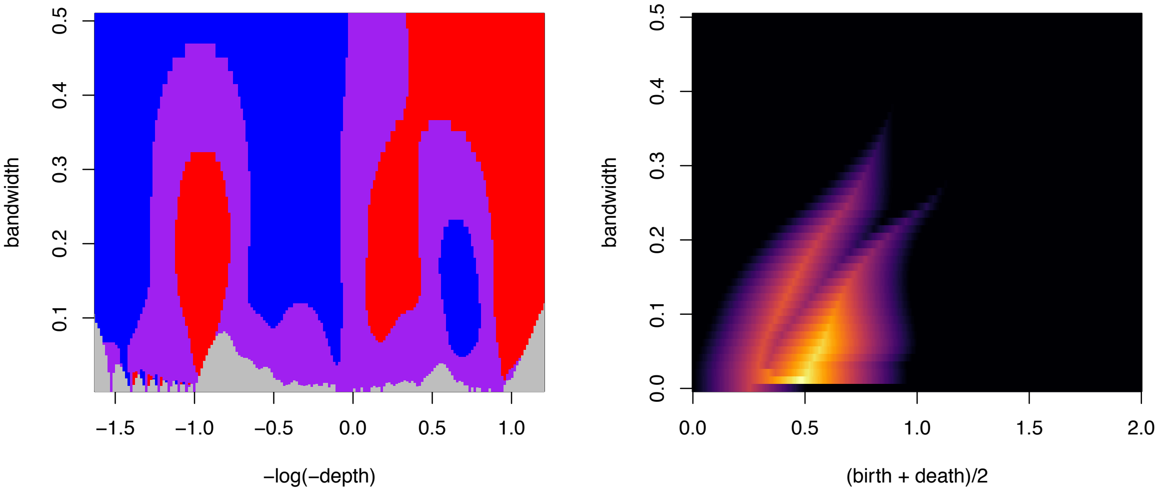

Figure 2.10: `SiZer`, the 1^{st} and 2^{nd} Persistence Flamelets of dimension 0. In order to facilitate the comparison with `SiZer`, the Persistence Flamelets is projected and represented as a matrix.

In this case we can see that the two peaks appearing in the 2^{nd} Persistence Flamelets correspond to the two points in the diagram (which in turn correspond to the two bumps we can see in the KDE in Figure 2.10). As we can see from Figure 2.8, the 2^{nd} Persistence Flamelets behaves differently than 1^{st} Persistence Flamelets; when the bandwidth grows in fact, the two secondary peaks are smoothed away.

Figure 2.10 shows the comparison with `SiZer`, and it is easy to see that the two approaches lead to very similar conclusions. The three peaks appear for $h = 0.05$, then one of them disappear at around $h = 0.25$, one other around $h = 0.35$ and, the last one always survives (in the given range of bandwidths).

Earthquakes II / Locations For our second example we consider earthquake data coming from the `USG catalog`. Our sample consists of the locations, expressed in latitude and longitude, of 6500 events with magnitude higher than 5, taking place between June 2013 and June 2017. The 2–dimensional density p generating the data $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$[3] can still be estimated using the kernel density estimator with a Gaussian Kernel:

$$\begin{aligned} \widehat{p}(\boldsymbol{x}) &= \frac{1}{n}\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x} - \boldsymbol{X}_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2\pi|\boldsymbol{H}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{X}_i)^t \boldsymbol{H}^{-1}(\boldsymbol{x} - \boldsymbol{X}_i)\right\}. \end{aligned}$$

Notice that in the multivariate case, the bandwidth is not a scalar but rather a matrix $\boldsymbol{H}$, however we chose an isotropic Gaussian Kernel, which corresponds to imposing a spherical structure to the covariance matrix

$$\boldsymbol{H} = h\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad h \in \mathbb{R}^+,$$

so that the kernel density estimator expression can be simplified as follows:

$$\widehat{p}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2\pi h} \exp\left\{-\frac{1}{2h^2}(\boldsymbol{x} - \boldsymbol{X}_i)^t(\boldsymbol{x} - \boldsymbol{X}_i)\right\}.$$

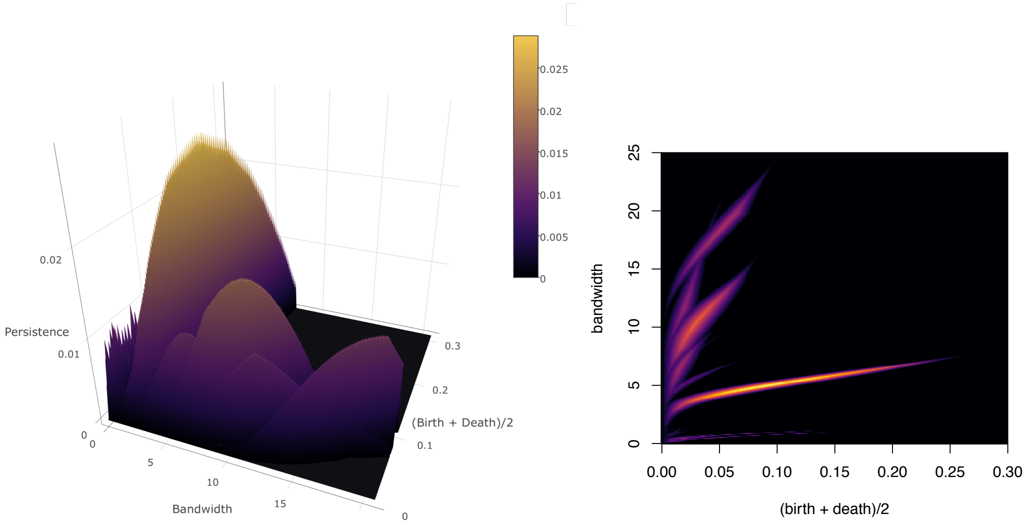

Figure 2.11: Dimension 1 Persistence Flamelets for earthquakes locations KDE (left) and its projection (right).

[3] Since we are trying to highlight the difference between univariate and multivariate densities, (only) in this section we will use the bold notation for vectors.

Earthquakes are concentrated around circular structures, also known as *plates*. According to Plate Tectonics, in fact, the Earth's lithosphere is broken into 7 main plates, plus a number of minor ones. Since earthquakes are caused by the movements of neighboring plates, the density p naturally inherits the Earth's plates structure. In terms of topology, plates can be thought of loops, or dimension 1 Homology Groups.

The dimension 1 Persistence Flamelets of the kernel density estimator $\widehat{p}$ can be employed to assess whether or not kernel density estimators are able to recover these loops. The Persistence Flamelets shown in Figure 2.11 presents 7 crests, each of them representing one persistent loop in $\mathcal{F}$; this seems to suggest that at, different resolution, the kernel density estimator is able to recover all the 7 main plates. Notice that as opposed to the $0^{\texttt{th}}$ dimensional case, where there is always one feature, the global maximum, dominating all the others, when analysing loops we can limit our analysis to the $1^{\texttt{st}}$ Persistence Flamelets. In this example specifically, the Persistence Flamelets shows that there is one loop that persists noticeably more than all the others; as persistence is a measure of the importance of a feature, this suggests that there is one plate which is more neatly detected than all others. This is represent the contour of the Philippine plate, which is not surprising, since more than 26% of the seismic activity in the given time interval was concentrated in the area between Philippine and Japan.

2.5.2 Bandwidth Selection

As the Persistence Flamelets is defined as the topological summary of a scale space, it is immediate to see that it can be exploited in the *exploration approach*, to asses the impact of the level of smoothness. However, as picking an optimal level of smoothing can be though of as a way of assessing whether or not a feature in a smooth is relevant, it may also play a role in the context of *bandwidth selection*, and it can be used to choose a "topologically–aware" bandwidth.

If evaluating the importance of higher dimensional topological features such as loops or voids is challenging from the point of view of exploration, this is even more true for the selection approach, where the topological structure is usually ignored (with the exception of local modes as in Genovese et al. (2016)). More critically, standard approaches for such as cross validation methods have proven to fail when the density is singular, i.e. concentrated around lower dimensional structures (Genovese et al. 2016).

Intuitively, since persistence can be interpreted as a measure of the importance of each feature, bandwidths corresponding to peaks in the Persistence Flamelets result in estimators that highlight the most prominent features in the density. By selecting the value of h that maximise the Persistence Flamelets, the *topologically–aware* $\widehat{h}_{\mathrm{TA}}$, we are forcing the density estimator to retain the most relevant topological treats.

Let us consider again the **Earthquake II** example. By choosing the value of h that maximise the Persistence Flamelets, we are forcing the density estimator to emphasize the most persistent loop. The kernel density estimator $\widehat{p}_{h_{\mathrm{TA}}}$, shown in Figure 2.12, is in fact concentrated around the Philippine plate, as we could expect.

To understand why such a topologically–aware bandwidth selection heuristic may be

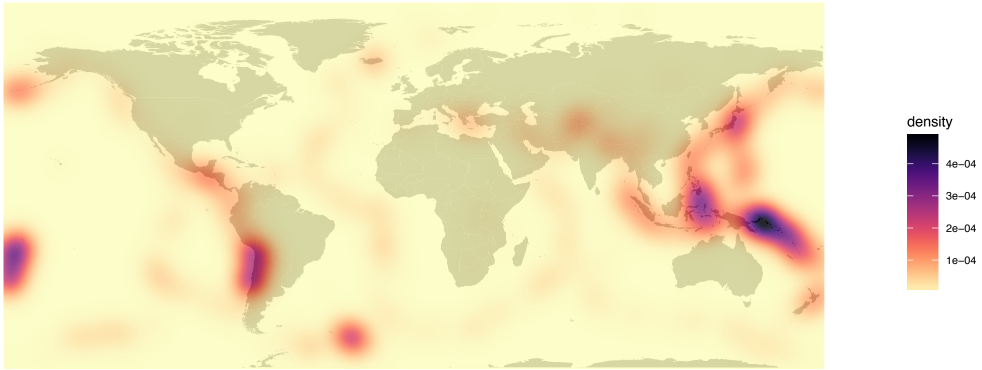

Figure 2.12: Density estimation with the topologically aware bandwidth $\widehat{h}_{\mathrm{TA}}$.

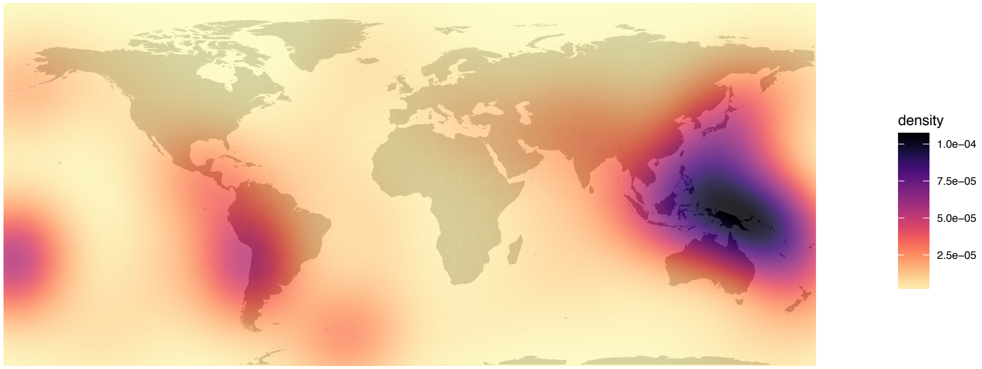

Figure 2.13: Density estimation with extended Silverman Normal bandwidth $\widehat{h}_{\mathrm{S}}$.

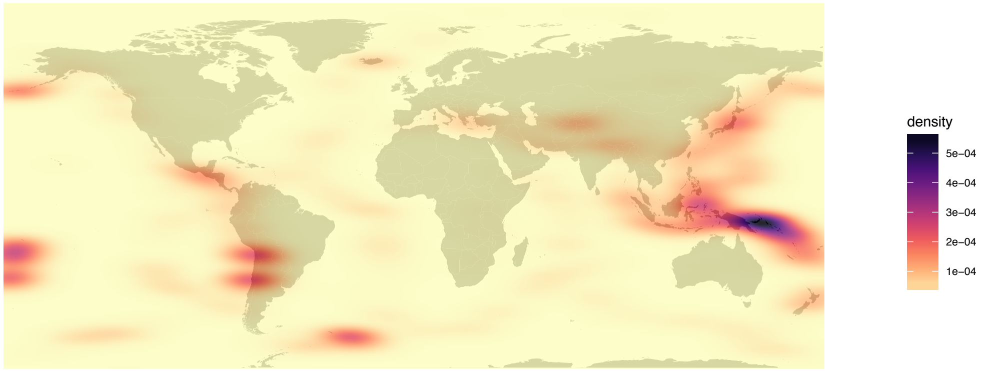

Figure 2.14: Density estimation with anisotropic Plug–in bandwidth matrix $\widehat{\boldsymbol{H}}_{\mathrm{PI}}$.

useful, let us compare it with more established methods for bandwidth selection: Silverman's Normal Rule and a Plug–in bandwidth selection criterion. We intentionally ignore cross validation methods because, as we stated in the previous section, they show poor behaviour in this setting.

The first alternative we consider is an extension of Silverman Normal Rule, one of the most famous "rule of thumb" for bandwidth selection, to the case of densities with singular features, as detailed in Genovese et al. (2017) and Chacón, Duong, and Wand (2011). More specifically, given a sample $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\} \in \mathbb{R}^D$, from some distribution P, the optimal bandwidth h for recovering the d–dimensional features is

$$\widehat{h}_{\mathrm{S}} = \left(\frac{4}{n(d+2)}\right)^{\frac{2}{4+d}} s,$$

where $s = D^{-1}\sum_{j=i}^{D} s_j^2$ and s_j^2 is the variance of the $j^{\texttt{th}}$ variable. Despite the fact that we set $d = 1$, in order to take into account the loop structure, the density estimator, shown in Figure 2.13, does not seem to recover any of the plates at all.

The second approach we consider is a Plug–in bandwidth estimator $\widehat{\boldsymbol{H}}_{\mathrm{PI}}$, obtained by minimizing the AMISE (Asymptotic Mean Integrated Square Error) w.r.t. the bandwidth h; details are given in Chacón, Duong, and Wand (2011)}. Since limiting the case of scalar bandwidths, as we did until here, may seem too restrictive, in this final example we relax the hypothesis of spherical covariance and do not impose any structure on the bandwidth matrix $\boldsymbol{H}$. The additional complexity of the estimator does not however result in a better estimation: as we can see in Figure 2.14, the plates structure of the true density is still not recognizable.

Chapter 3

Topological Supervised Learning

3.1 Topological Kernels

Persistence diagrams have several drawbacks that have limited their popularity in statistical inference. For example, a collection of Persistence Diagrams $\{D_1, \dots, D_n\}$, does not have a unique mean (Turner et al. 2014); perhaps even more critically, despite the fact that $\mathcal{D}$ is a Polish space and that the existence of a probability distribution on it has been proved by Mileyko, Mukherjee, and Harer (2011), it is still not clear how to derive it. In general, the metric structure of the space of persistence diagrams may not be rich enough for statistical learning.

We approach supervised learning with Persistence Diagrams as covariates by translating Persistence Diagram into inner product spaces using kernels. A kernel K on a space $\mathcal{X}$ is a symmetric binary function $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}^+$ that can roughly be interpreted as a measure of similarity between two elements of $\mathcal{X}$. Every kernel is associated to an inner product space (Scholkopf and Smola 2001); exploiting this correspondence, kernels allow to perform directly most statistical tasks such as classification (Cristianini and Shawe-Taylor 2000), regression (Härdle 1990), or testing (Gretton et al. 2012), without explicitly computing, or explicitly knowing, the probability distribution that generated the observations.

One popular family of kernels for a geodesic metric space $(\mathbb{X}, d)$ is the *exponential kernel*

$$k(x, y) = \exp\{-d(x, y)^p / h\} \qquad p, h > 0$$

where $h > 0$ is the bandwidth parameter; for $p = 1$ this is the Laplacian kernel and for $p = 2$ this is the Gaussian kernel. As the space of Persistence Diagrams is a Geodesic Space, it is possible to use this class to define a *Topological kernel* to be deployed in statistical learning.

Definition 3.1 (Geodesic Topological Kernel)**.** Let $\mathcal{D}$ be the space of persistence diagrams, and let $h > 0$, then the Geodesic Gaussian Topological (GGT) kernel $K_{\mathrm{GG}} : \mathcal{D} \times \mathcal{D} \mapsto \mathbb{R}^+$ is defined as

$$K_{\mathrm{GG}}(D, D') = \exp\left\{-\frac{1}{h} W_{L^2,2}(D, D')^2\right\} \qquad \forall\, D, D' \in \mathcal{D}.$$

Analogously, the Geodesic Laplacian Topological Kernel (GLT), K_{GL} is defined as:

$$K_{\mathrm{GL}}(D, D') = \exp\left\{-\frac{1}{h} W_{L^2,2}(D, D')\right\} \qquad \forall\, D, D' \in \mathcal{D}.$$

It may seem natural to extend the properties of the standard (Euclidean) Gaussian and Laplacian kernels to their geodesic counterpart on $\mathcal{D}$, however, it turns out that the metric structure of the space $\mathcal{D}$ may introduce some limitations, especially with respect to positive definiteness; as shown in Feragen, Lauze, and Hauberg (2015), in fact, a Geodesic Gaussian kernel on a metric space is positive definite only if the space is flat.

Theorem 3.1 (Feragen et al.). *Let $(\mathbb{X}, d)$ be a geodesic metric space and assume that the Geodesic Gaussian kernel on $\mathbb{X}$ $k(x, y) = \exp\{-d^2(x, y)/h\}$ is positive definite for all $h > 0$. Then $(\mathbb{X}, d)$ is flat in the sense of Alexandrov (see Bridson (1999) for more information).*

This is not the case for the space of Persistence Diagram, which has been proved to be curved by Turner et al. (2014). We say that a geodesic metric space is CAT(k) if its curvature is bounded from above by k.

Theorem 3.2 (Turner et al.). *The space of persistence diagrams $\mathcal{D}$ with $W_{L^2,2}$ is not CAT(k) for any $k > 0$, and it is a non–negatively curved Alexandrov space.*

We can now characterize the Geodesic Gaussian Kernel.

Lemma 3.1. *The Geodesic Gaussian Kernel on $\mathcal{D}$ is not positive definite.*

The proof is a trivial consequence of Theorem 3.1 and Theorem 3.2. Characterizing the Geodesic Laplacian kernel is not as easy, although it has shown empirically to be indefinite as well (Reininghaus et al. 2015).

3.1.1 The competition

This is not the only, nor the first, attempt to transform persistence diagrams into a more "inferential–friendly" object. Previous works in this direction however followed a different strategy and tackled the problem by explicitly deriving a feature map $\Phi : \mathcal{D} \mapsto \mathcal{H}$ from persistence diagrams to some Hilbert space $\mathcal{H}$. The link between this and our approach is that any feature map Φ corresponds to a kernel K (Cristianini and Shawe-Taylor 2000, Scholkopf and Smola (2001)) defined as $K(D, D') = \langle \Phi(D), \Phi(D') \rangle_{\mathcal{H}}$, for every $D, D' \in \mathcal{D}$.

We briefly review the two main families of feature maps Φ: 1. feature maps derived from the Triangle function and 2. feature maps derived from the Dirac Delta function. A common element to the methods presented in the following is that the embedding is defined point–wise, for each element of the persistence diagram, at first. The structure of the diagram must be later recovered as a summary, whereas the geodesic kernel maintains it directly, as it always consider the persistence diagram as a whole.

Triangle Function As we have already seen in Section 2.3, the first way of translating each point $x \in D$ into a space of function is through the triangle function $T_z(y)$ defined in Equation (2.1), which allows to represent a persistence diagram as a collection of piecewise linear functions; for any $k \in \mathbb{N}^+$, Persistence Landscapes $\lambda_D^k(y)$ are defined by taking the $k^{\mathtt{th}}$ outermost line of the collection. It immediately follows that for any given $k \in \mathbb{N}^+$ the feature map $\Phi(D)$ is defined as $t \mapsto \lambda_D(k, t)$, meaning that it is possible to define a kernel from the Persistence Landscape $K_\lambda(D, D')$ (and analogously for the Silhouette), but since in practice it has shown poor performances (as shown in Reininghaus et al. (2015)), these tools are typically used as they are or summarized in some other way.

Dirac Delta Functions The second way of mapping each $x \in D$ to a space of function is through Dirac delta functions δ_x. Every Persistence diagram D can be uniquely represented as the sum of Dirac delta functions δ_x, one for each $x \in D$; since δ_x are defined in a Hilbert space, their sum will as well.

Reininghaus et al. (2015) use this representation as initial condition for a heat diffusion problem, and define a new feature map $\Phi(D)$ as

$$t \mapsto \frac{1}{4\pi\sigma} \sum_{x \in D} \mathsf{e}^{-\frac{\|t-x\|^2}{4\sigma}} - \mathsf{e}^{-\frac{\|t-\bar{x}\|^2}{4\sigma}},$$

where if $x = (b, d)$ then $\bar{x} = (d, b)$. The feature map $\Phi(D)$ defines the Persistence Scale Space kernel K_{PSS}:

$$K_{\text{PSS}}(D, D') = \frac{1}{8\pi\sigma} \sum_{x \in D} \sum_{y \in D'} \mathsf{e}^{-\frac{\|x-y\|^2}{8\sigma}} - \mathsf{e}^{-\frac{\|x-\bar{y}\|^2}{8\sigma}} \qquad \forall\, D, D' \in \mathcal{D},$$

which is the most similar in spirit to the Geodesic Kernels. K_{PSS} is a heat kernel, and is stable with respect to $W_{L^\infty,1}$.

Another kernel built from Dirac Delta functions is the Persistence Weighted Gaussian Kernel (Kusano, Hiraoka, and Fukumizu 2016), defined as

$$K_{\text{PWG}}(D, D') = \exp\left(-\frac{d_{\text{G}}(D, D')^2}{2\,\sigma^2}\right)$$

where

$$\begin{aligned} d_{\text{G}}(D, D') = &\sum_{x \in D} \sum_{x' \in D} w_{\text{arc}}(x)\, w_{\text{arc}}(x')\, k_{\text{G}}(x, x') \\ &+ \sum_{y \in D'} \sum_{y' \in D'} w_{\text{arc}}(y)\, w_{\text{arc}}(y')\, k_{\text{G}}(y, y') \\ &- 2 \sum_{x \in D} \sum_{y \in D'} w_{\text{arc}}(x)\, w_{\text{arc}}(y)\, k_{\text{G}}(x, y), \end{aligned}$$

$$w_{\text{arc}}(x) = \arctan\left(C \cdot \text{pers}(x)^q\right),$$

and k_{G} is the Euclidean Gaussian kernel with variance τ. The Persistence Weighted Gaussian Kernel, much like the Persistence Silhouette, allows to explicitly control the effect of persistence. However, the choice of the different 4 tuning parameters (q, σ, τ, C) may be unfeasible in most real data applications.

The main difference with respect to our Geodesic Kernels is that K_{PSS}, K_{PWG} and even K_λ are positive definite by construction. Despite being indefinite, however, the Geodesic Kernels are a more sensible measure of similarity.

	D	D'	$D_\emptyset$
D	1.000	0.040	0.483
D'	0.040	1.000	0.006
$D_\emptyset$	0.483	0.006	1.000

Table 3.1: Geodesic Gaussian Kernel matrix for the three diagrams shown in Figure 3.1.

	D	D'	$D_\emptyset$
D	0.005	0.023	0.000
D'	0.023	0.119	0.000
$D_\emptyset$	0.000	0.000	0.000

Table 3.2: Persistence Scale Space Kernel matrix for the three diagrams shown in Figure 3.1.

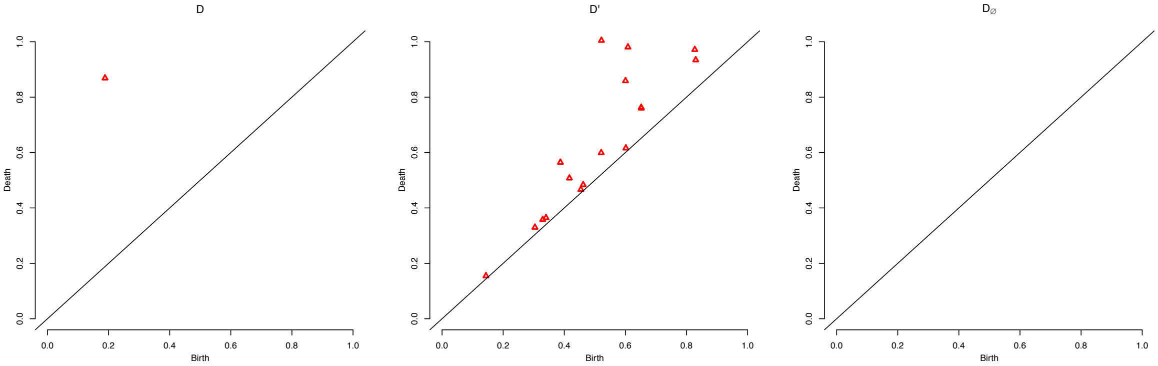

Figure 3.1: (From left to right) Three Persistence diagrams: D, D', $D_\emptyset$.

Let us examine the behavior of the kernels with respect to the empty diagram $D_\emptyset$ to make this more clear. This will be especially relevant later, when analyzing posturography data (see Section 3.3). Although not all diagrams are equally different from the empty diagram $D_\emptyset$, K_{PSS} and K_{PWG} do not capture this diversity as neatly as the Geodesic Kernels.

In the PSS approach, for example, $\Phi(D_\emptyset) = 0$ by definition. This results in $K_{\text{PSS}}(D_\emptyset, D) = \langle \Phi(D_\emptyset), \Phi(D) \rangle = 0$, for every $D \in \mathcal{D}$, including $D_\emptyset$ itself, leading to the paradoxical conclusion that $K_{\text{PSS}}(D_\emptyset, D_\emptyset) = 0$, as shown in Table 3.2.

The Geodesic Kernels, on the other hand, are built on the Wasserstein distance and since $W_{L,p}(D, D_\emptyset) \neq 0$ for any $D \neq D_\emptyset$, they retain more information, as can be seen in Table 3.1.

Although positive definiteness is a rather attractive quality in a kernel (Scholkopf and Smola 2001), the indefiniteness of our kernel does not affect its performances in supervised settings. Notice that we are not claiming that our kernel is superior in general, in fact due to their positive definiteness K_{PSS} and K_{PGW} can be used outside supervised learning, we are instead proposing an alternative that exploiting the predictive power of the negative part of the kernels can perform better in this class of problems. We now show some applications to real data to support our thesis.

3.2 Regression / Fullerenes

Buckyballs fullerenes are spherical pure carbon molecules artificially synthesized in the ‘70, then discovered in nature in the’90, which have recently gained much attention after C60 has being identified as the largest molecule detected in space (Berné and Tielens 2012). The typical trait of Buckyballs fullerenes is that atoms’ linkage can form either pentagons or hexagons, so that the configuration of the molecule resembles a soccer ball (hence the name). Our goal is to show that the topology of the molecule can be used directly to explain its Total Strain Energy (measured in Ev); given a sample $\{X_1, \dots, X_n\}$ of Fullerenes we model their Total Strain Energy, Y as a function of their Persistence Diagrams $\{D_1, \dots, D_n\}$:

$$Y_i = m(D_i) + \varepsilon_i \qquad \forall\, i \in \{1, \dots, n\},$$

where ε_i is the usual 0–mean random error.

As in standard nonparametric regression, we can estimate the regression function $m(\cdot)$ with the Nadaraya–Watson estimator (Härdle et al. 2012), defined as:

$$\widehat{m}(D) = \frac{\sum_{i=1}^{n} Y_i\, K(D, D_i)}{\sum_{i=1}^{n} K(D, D_i)},$$

where D is a generic persistence diagram. Since the kernel function K involved in the Nadaraya–Watson estimator, needs not be positive definite, we can use the Geodesic kernels to extend nonparametric regression to the case of persistence diagrams as covariate.

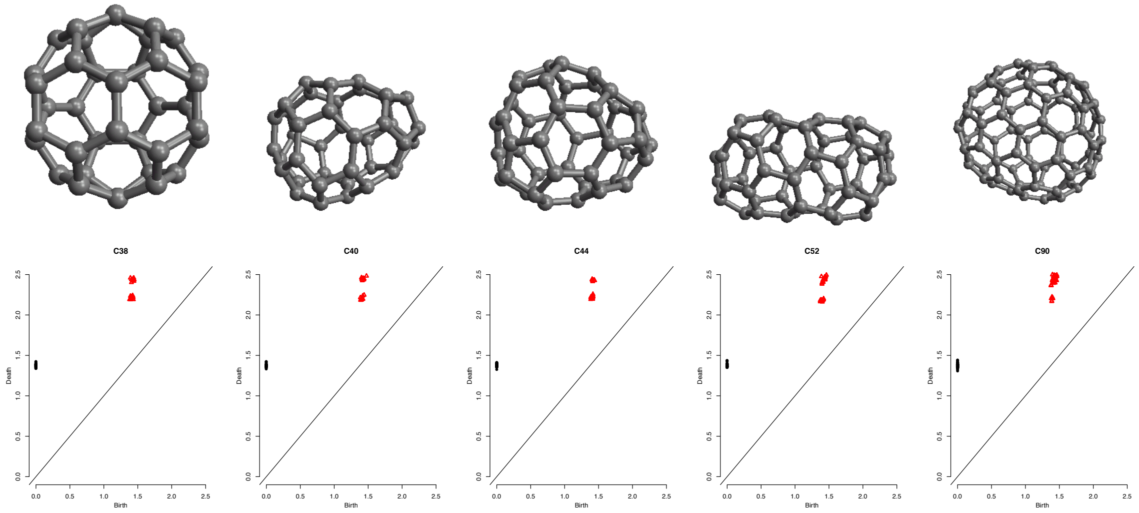

Figure 3.2: Topological configurations of some fullerenes (top) and corresponding persistence diagrams (bottom). From left to right: C38(C2v), C40(C1), C44(C1), C52(C2), C90(C1).

	C38	C40	C42	C44	C48	C52	C84	C86	C90	C100
n	17	40	45	89	79	96	24	19	46	80
$\bar{Y}$	27.50	28.29	28.46	29.12	31.21	32.59	29.34	29.88	31.29	34.41
$\widehat{\sigma}$	1.35	1.62	1.35	1.78	1.56	1.57	1.29	0.80	1.21	1.24

Table 3.3: Number of observations (n), mean ($\bar{Y}$) and standard deviation ($\widehat{\sigma}$) of TSE for each type of fullerenes in the sample.

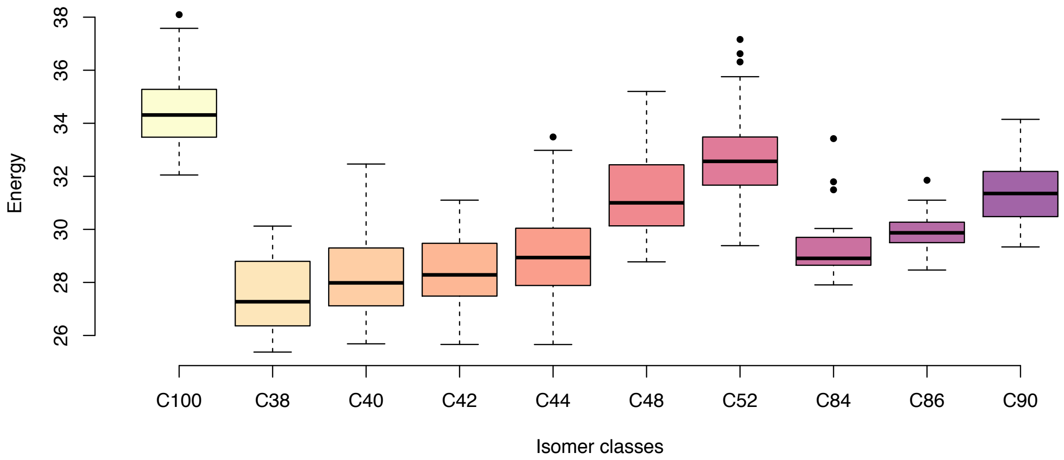

Figure 3.3: Energies for the 10 different classes of isomers. It is worth noticing that Fullerenes with higher numbers of atoms do not necessarily have higher energy.

We fit the model using data from $n = 535$ molecules of 10 different types of Fullerenes. The sample is unbalanced, as the number of configurations available for each Fullerene depends on the number of atoms composing it and advances in research (Table 3.3). For each molecule, the data (freely available at `http://www.nanotube.msu.edu/fullerene/fullerene-isomers.html` consists of the coordinates of the atoms taken from Yoshida's Fullerene Library and then re–optimized with a Dreiding–like forcefield. We carry our analysis using both the `R` package `TDA` (Fasy et al. 2014) and the `C++` library it refers to `Dionysus` (Morozov 2012).

Since there is no clear pattern for connected components and, as we could expect, there is only one relevant void for each molecule, we decided to focus on features of dimension 1, which seem to be the most informative. As we can see from Figure 3.2, loops in the diagrams are, in fact, clearly clustered around two centers, which represent the pentagons and the hexagons formed by the carbon atoms. Interestingly enough, the Wasserstein distance and, hence, both the geodesic kernels, fully recover the class structure induced by the isomers, as we can see in Figure 3.4.

	Geodesic Gaussian Kernel	Geodesic Laplacian Kernel
Nonparametric regression	339.89	342.14
Semiparametric regression	1049.02	331.04

Table 3.4: Residual Sum of Squares.

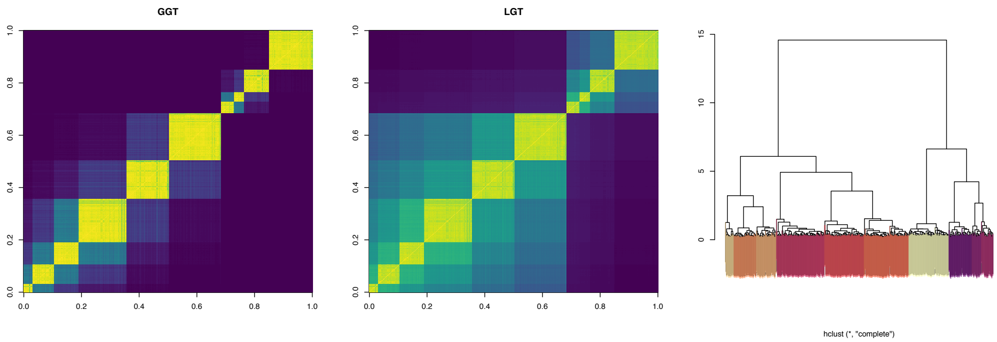

Figure 3.4: Kernel Matrix for the Geodesic Gaussian Kernel (left), Geodesic Laplacian Kernel (center), Hierarchical Clustering built from the Wasserstein distance with complete linkage (right). Colors represent the different isomer classes as shown in Figure 3.3.

We estimate the regression function $m(D)$ using both the Laplacian and the Gaussian geodesic kernels; the estimator resulting from the GGT kernel is

$$\widehat{m}_{\mathrm{GG}}(D) = \frac{\sum_{i=1}^{n} Y_i \exp\left\{-\frac{1}{h} W_{L_\infty,2}(D, D_i)^2\right\}}{\sum_{i=1}^{n} \exp\left\{-\frac{1}{h} W_{L_\infty,2}(D, D_i)^2\right\}} \qquad \forall\, D \in \mathcal{D};$$

analogously for the LGT kernel. Moreover, in order to take into account the group structure naturally induced by the isomers, we considered a model with a fixed group intercept, i.e:

$$Y_{ij} = \alpha_j + m(D_{ij}) + \varepsilon_{ij},$$

where D_{ij} denotes the persistence diagram of the $i^{\mathtt{th}}$ isomer of the $j^{\mathtt{th}}$ molecule. We fit the resulting partially linear model using Robinson's trimmed estimator, as detailed in Li and Racine (2007).

After choosing the bandwidth h via Leave–One–Out cross validation, we compare the different models in terms of Residual Sum of Squares (RSS). As we can see from Table 3.4, the two kernels yield similar results when used in a fully nonparametric estimator, while the Laplacian kernel performs better when adding the group intercept to the model. This can be understood by looking at the kernel matrices (Figure 3.4); the Gaussian Kernel has a sharper block structure than the Laplace Kernel, which makes it better at discriminating the 10 molecule classes. However, when the group structure is taken into account by the model itself, this clustered structure leads to worse prediction.

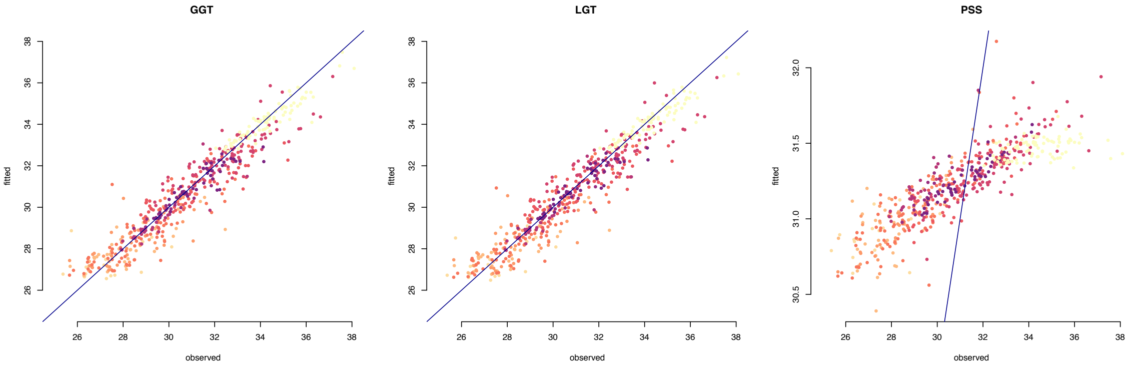

Figure 3.5: Observed vs fitted plot for the fully nonparametric model fitted with the Geodesic Gaussian (left), Geodesic Laplacian (center) and the Persistence Scale Space kernel (right). Colors represent the different isomer classes as shown in Figure 3.3.

Finally, we compare the performance of our geodesic kernels with the Persistence Scale Space kernel K_{PSS} by using the same data to fit

$$\widehat{m}_{\text{PSS}}(D) = \frac{\sum_{i=1}^{n} Y_i \, K_{\text{PSS}}(D, D_i)}{\sum_{i=1}^{n} K_{\text{PSS}}(D, D_i)}.$$

As we can clearly see from the fitted-vs-observed plots in Figure 3.5, the positive definiteness of the PSS kernel does not result in more accurate prediction, as both K_{GG} and K_{LG} outperform it.

3.3 Classification / Posturography

For our second example we analyze data from a posturography experiment available at `https://physionet.org/physiobank/database/hbedb/`. Subjects standing on a platform were asked to close their eyes and stand still for some time. Researcher then recorded the center of pressure on the platform over a period of 60 seconds; details are available in Santos and Duarte (2016). In order to characterize the oscillation's pattern using `TDA`, we build a Persistence Diagram for each of the 320 traces, 160 of which were recorded on a rigid platform, and 160 on a soft one.

We focus on dimension 1 topological features. Intuitively, in fact, a loss of equilibrium results in sudden movements, which generate cyclical structures; we can consider the number and the persistence of loops as a measure of the signal's variability. Figure 3.6 shows one trajectory for each of the conditions. Data coming from the rigid platform do not present any loop at all, causing the diagram to be empty (as we are only considering dimension 1 features).

Although not all observations are quite as well distinguishable, it is generally true that subjects standing on the rigid platform are more stable and their persistence diagrams are more likely to be empty.

This kind of data fits perfectly in the TDA framework, as coordinates, which in this case represent the direction and the time of the loss of equilibrium, are not relevant to our problem and may be misinterpreted.

We show how the Persistence Diagram of a trace can be used to infer whether each trajectory was recorded on a soft or a rigid platform. This is a binary classification problem, which we solve using the Geodesic Kernels. Standard kernel–based classifiers such as Support Vector Machines require a positive definite kernel, we thus consider an extension to SVM for indefinite kernels proposed by Loosli, Canu, and Ong (2016), KSVM. Details are given in Appendix C.

As we can see from Table 3.5, the accuracy of the classification is far superior when using KSVM with the Geodesic Gaussian Kernel K_{GG} (and results are identical for K_{LG}) rather than the standard SVM with the positive definite K_{PSS}, and this result is not surprising because several of the diagrams corresponding to trajectories on the Rigid Platform are empty. Although there are algorithms, such as KSVM and others (Luss and d'Aspremont 2008),

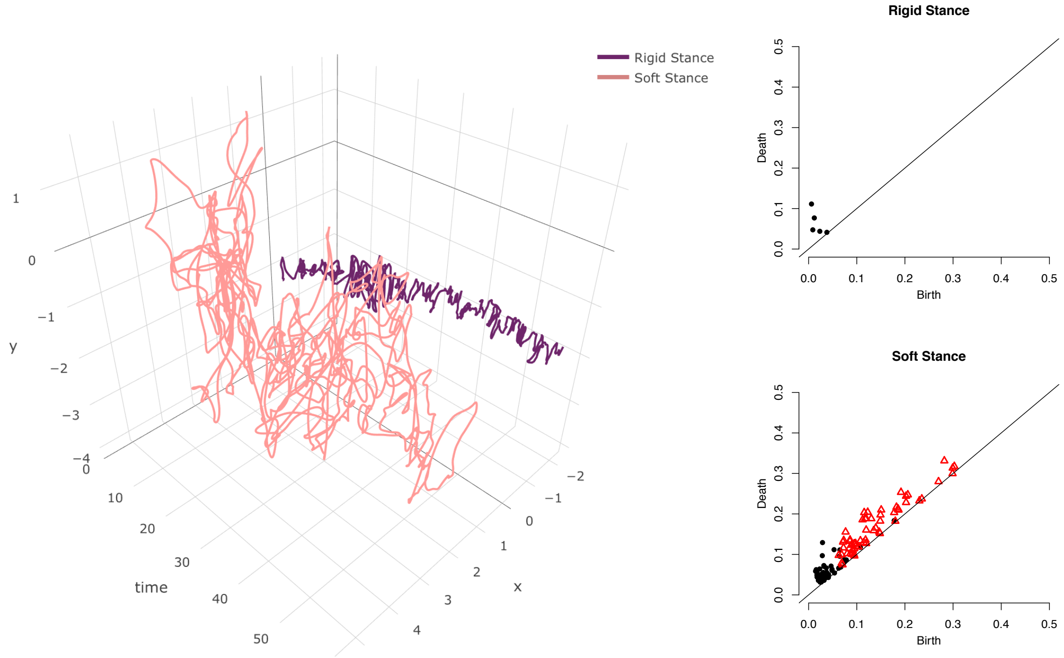

Figure 3.6: Trajectory of a subject standing on a soft platform (in pink) and on a rigid one (in purple). On the left, the corresponding persistence diagrams.

	KSVM	PSS–SVM	Clip	Flip	Square
Mean	2.82%	3.31%	2.84%	2.86%	2.87%
Standard Deviation	0.087	0.198	0.159	0.141	0.166

Table 3.5: Average Misclassification Rate for the 10–fold Cross Validation and corresponding variance.

designed to explicitly solve the SVM optimization problem when the kernel is indefinite, a very common way to deal with indefinite kernels K is to just substitute the kernel matrix $\mathbb{K}$, whose $(i,j)^{\mathtt{th}}$ entry is defined as $\mathbb{K}_{ij} = K(D_i, D_j)$, with some positive definite approximation of it. Denote by $\mathbb{K} = U\,\Lambda\,U^{\mathtt{t}}$ the spectral decomposition of the indefinite matrix $\mathbb{K}$, where U is an orthogonal matrix and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is the diagonal matrix of (real by symmetry) eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$. We consider the following heuristics to obtain a positive definite kernel matrix $\widetilde{\mathbb{K}}$:

- **clip:** set to 0 negative eigenvalues of $\mathbb{K}$; that is, $\widetilde{\mathbb{K}}_{\mathtt{c}} = U\,\widetilde{\Lambda}_{\mathtt{c}}\,U^{\mathtt{t}}$ where
$$\widetilde{\Lambda}_{\mathtt{c}} = \mathrm{diag}(\,\max(\lambda_1, 0), \ldots, \max(\lambda_n, 0));$$
- **flip:** take the absolute value of the eigenvalues of $\mathbb{K}$; that is, $\widetilde{\mathbb{K}}_{\mathtt{f}} = U\,\widetilde{\Lambda}_{\mathtt{f}}\,U^{\mathtt{t}}$ where
$$\widetilde{\Lambda}_{\mathtt{f}} = \mathrm{diag}(|\lambda_1|, \ldots, |\lambda_n|);$$
- **square:** square the eigenvalues of $\mathbb{K}$; that is, $\widetilde{\mathbb{K}}_{\mathtt{s}} = U\,\widetilde{\Lambda}_{\mathtt{s}}\,U^{\mathtt{t}}$ where
$$\widetilde{\Lambda}_{\mathtt{s}} = \mathrm{diag}(\lambda_1^2, \ldots, \lambda_n^2).$$

We compare the performance of KSVM with that of a standard SVM trained on $\widetilde{\mathbb{K}}$. The three heuristics we consider in order a positive definite version of the kernel matrix $\mathbb{K}$. Results in Table 3.5 are rather reassuring, since they suggest that the good performance of the KSVM with K_{GG} it does not depend on the complexity of the specific solver, but rather on the discriminative power of the Geodesic Kernels themselves.

3.4 Topological Determinants of Brain Activity

Having so far analysed quantile and topological separately, we now propose a joint application of the previously introduced techniques, focusing on brain imaging data. In addition to providing an interesting question per se, as there is an obvious fascination with the quest for insights on how human minds work, their complexity encourages the development of new statistical tools, making neuroimaging a statistical goldmine in the last decades. Brain imaging exhibits in fact complex temporal and spatial dependency, which is non trivial to assess, especially because of the rather complex and still not entirely known geometry of the brain.

On top of the modelling challenges that naturally arises when studying such a complex object, neuroimaging data present the additional obstacle of being extremely high dimensional. We do not by any means claim to provide a complete analysis of such a complicated problem, but we consider section as a proof of concept instead. Our goal is twofold: 1. to show that there is something to be gained by adopting Topological Tools in the analysis of fMRI data, 2. provide a pipeline for recasting topological summaries in the framework of parametric modeling.

The first studies revolved around *Structural* imaging tools such as CT (Computer axial Tomography) or PET (Positron Emission Tomography). However, in addition to being potentially harmful to the patient due to the radiation involved, these tools only provide a static characterization of the brain and may not be used to investigate its activity, which is why in recent years, the interest shifted towards less invasive *Functional* imaging tools, most noticeably functional Magnetic Resonance Imaging (fMRI), ElectroEncephaloGram (EEG) or Diffusion Tensor Imaging (DTI) which are intended to capture the brain "in action". In this example we try to exploit the tools introduced in the previous sections to investigate brain activity, hence we turn to this second class of neuroimaging data, focusing on fMRI in particular.

The principle behind fMRI is that brain activation can be detected by analyzing the changes in blood oxygenation and blood flow corresponding to some task. fMRI data consists of collection of 3-dimensional magnetic resonance images acquired on a tight time grid. Each pixel, also called voxel, represent the intensity of the nuclear spin density, which is strictly related to the blood oxygenation and flow. More precisely, the most popular approach for performing fMRI is based on the Blood Oxygenation Level Dependent (BOLD) contrast, which allows to study the hemodynamic response to neural firing.

The biological justification for the BOLD is that the metabolic demand for oxygen and nutrients increases in relation to increases in neural activity increases, only in the affected regions of the brain. The Neural firing thus signals the extraction of oxygen from hemoglobin in the blood, which can be detected by the BOLD signal. It is worth highlighting that the BOLD signal is not a direct quantification of neural activation, as there may be changes in neural activity that do not necessarily change the metabolic demand of the region, hence they may not result in an increased need of hemoglobin. Despite being rather informative, the BOLD signal thus captures only partially the effect of a change of the neuronal activities corresponding to a task.

Statistical analysis of fMRI scans were initially aimed at investigating how the brain reacted to a specific stimulus, hence they drew on fMRI recorded during the execution of a given task. In recent years and despite initial skepticism, resting state fMRI, that is fMRI recorded while the patient is at rest, have started gaining momentum in the neuroimaging community (Van Dijk et al. 2009; Biswal et al. 2010), as they provide insights on deeper forms of brain activity.

3.4.1 Why Topology

The most common framework for investigate and model brain activity from a *functional* rather than *structural* standpoint is to represent the brain as a graph. In this setting, different brain areas, taken to be the nodes of the network, are connected by an edge when they show a similar behavior. The main advantage in adopting this approach is that brain activity can be analysed by exploiting tools from network analysis. Connectivity, hence topological characterization, is a core notion in this class of methods, as connected components of brain network consist of areas of the brain that show a similar behavior, regardless of their spatial proximity, and it is typically analysed by assessing network properties (e.g. small-world, scale free connectivity). We refer to Bullmore and Sporns (2009) Lee, Smyser, and Shimony (2013) and references therein for a general review of comparisons made using network summary statistics.

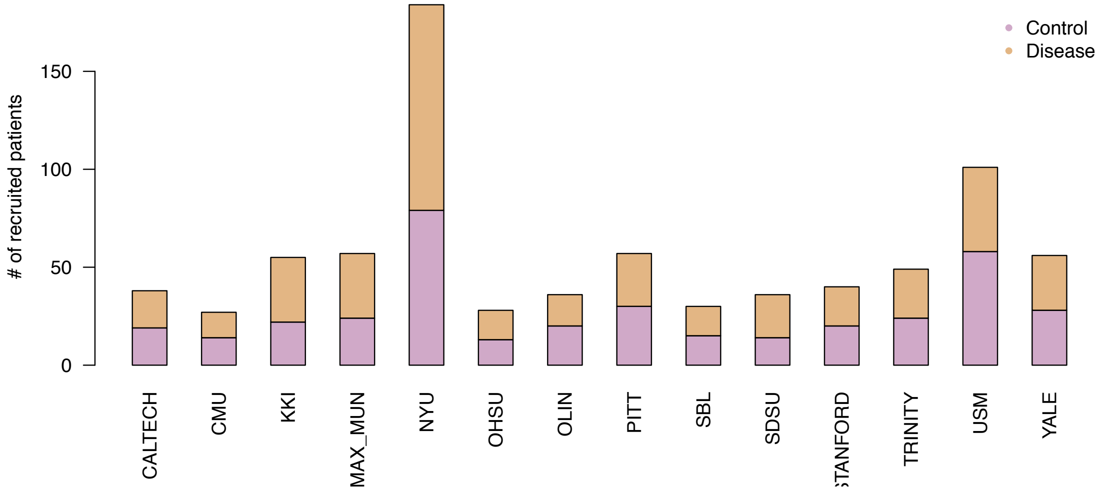

Figure 3.7: Data collection centers.

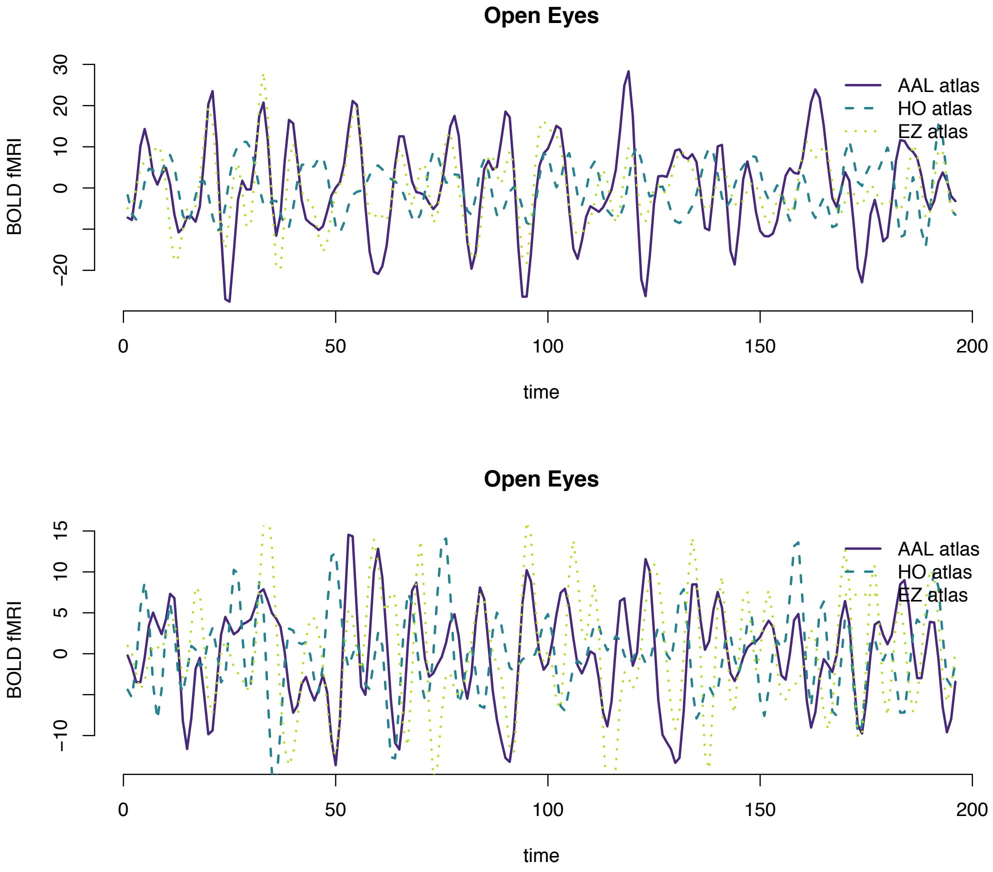

Figure 3.8: Raw time series for one ROI of each atlas, with open (top) and closed eyes (bottom).

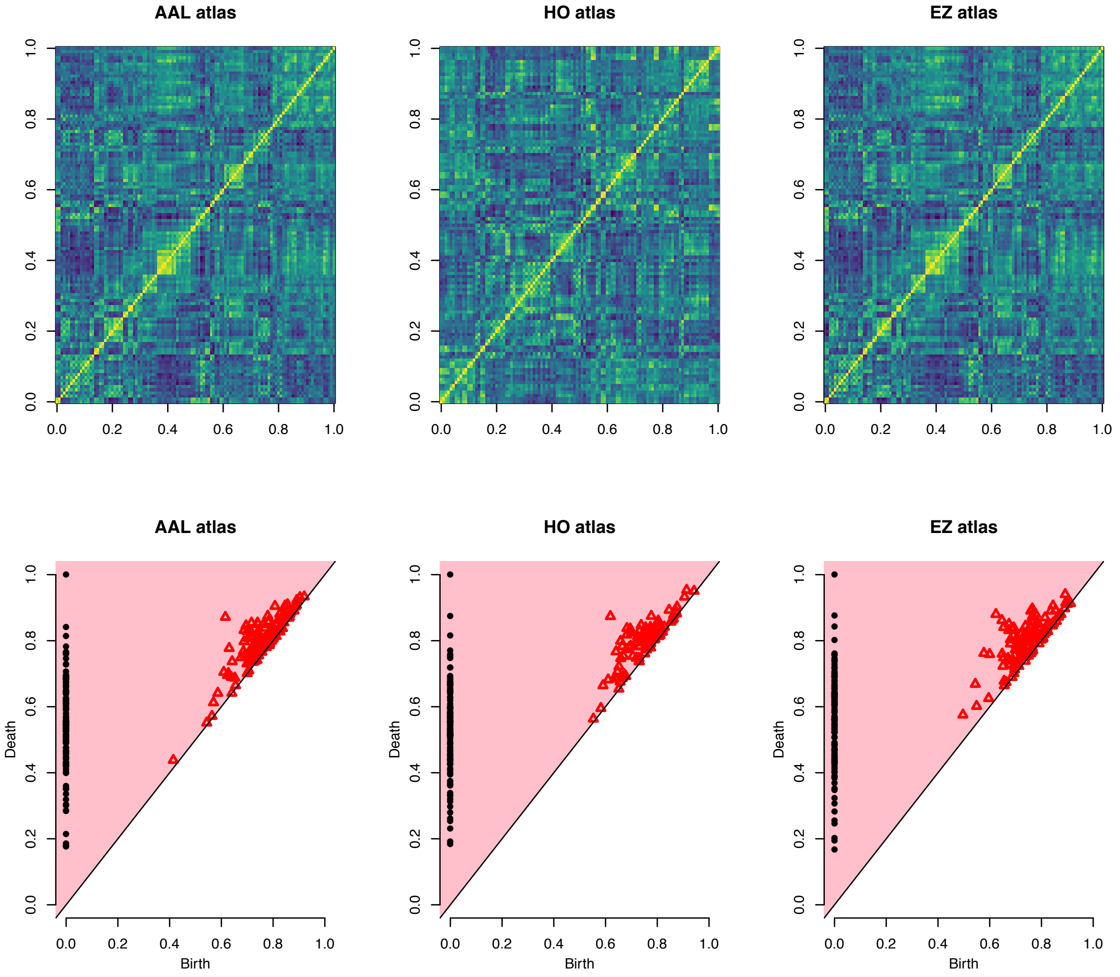

Figure 3.9: Correlation matrices (top) and corresponding Persistence Diagram (bottom) for one subject of the study with open eyes, with the area of the diagram highlighted in pink representing the confidence band around the diagonal.

Since building the brain network typically requires inducing some sparsity in the association structure of brain regions, as a fully connected graph would be difficult to interpret, techniques such graphical lasso or something are deployed to threshold the dependency measure. Our first aim is thus to assess whether the additional information provided by the persistence approach, that does not require any arbitrary thresholding choice, is relevant by trying to exploit Persistence Diagram for characterizing the brain activity.

Region of Interest The first challenge faced in building persistence diagrams on fMRI is the definition of "brain areas", or unit of observation.

Considering individual voxel of fMRI scans as unit of analysis has the obvious advantage of preserving all the information contained in the data, however, as the number of voxel in a scan can be of the order of millions, treating voxel directly can be computationally prohibitive. Moreover much of the information contained in the fMRI scan is redundant, as close voxel typically exhibit almost identical behavior, therefore neighboring voxel are typically aggregated into Region of Interest (ROI). One way to do so is to exploit fully data-driven tools such as Principal Component Analysis (Andersen, Gash, and Avison 1999) or Independent Component Analysis (Calhoun et al. 2001). This dimensionality reduction techniques identify the ROIs as lower-dimensional structures in the data, but are not easily interpretable. Regions defined following this approach may vary according to the activity performed during the scan and, even more critically, according to the subject performing it. Making comparison difficult, this kind of ROI may be more indicated for studies aimed at investigating subject specific behaviors rather than modelling multiple individuals, as in our case.

In order to make result comparable across subjects, we use a parcellation in Region of Interest defined by practitioners from anatomical proximity. The use of anatomical atlases allows an appealing interpretation of the results, since a brain division that corresponds to criteria defined by experts in the field should be more coherent with the state-of-the-art knowledge. Furthermore, a standardized parcellation has the additional benefit to permit comparison of results from different methodologies and different datasets. As results may depend on the specific atlas considered, we perform our analysis on multiple atlases and compare our findings. More specifically we compare the following three parcellations.

- **Automated Anatomical Labeling (AAL)** Adopted in more than three-quarters of the publications on functional brain networks, the Automated Anatomical Labeling (AAL) is the most popular anatomical template for fMRI. The AAL atlas distributed with the AAL Toolbox was fractionated to functional resolution (3x3x3 mm3) using nearest-neighbor interpolation, resulting in 116 regions.

- **Eickhoff-Zilles (EZ)** The EZ atlas was derived from the max-propagation atlas distributed with the SPM Anatomy Toolbox. The atlas was transformed into template space using the Colin 27 template (also distributed with the toolbox) as an intermediary and fractionated into functional resolution using nearest-neighbor interpolation.

- **Harvard Oxford (HO)** The HO atlas distributed with FSL is split into cortical and subcortical probabilistic atlases. A 25% threshold was applied to each of these atlases and they were subsequently bisected into left and right hemispheres at the midline (x=0). ROIs representing left/right WM, left/right GM, left/right CSF and brainstem were removed from the subcortical atlas. The subcortical and cortical ROIs were combined and then fractionated into functional resolution using nearest-neighbor interpolation.

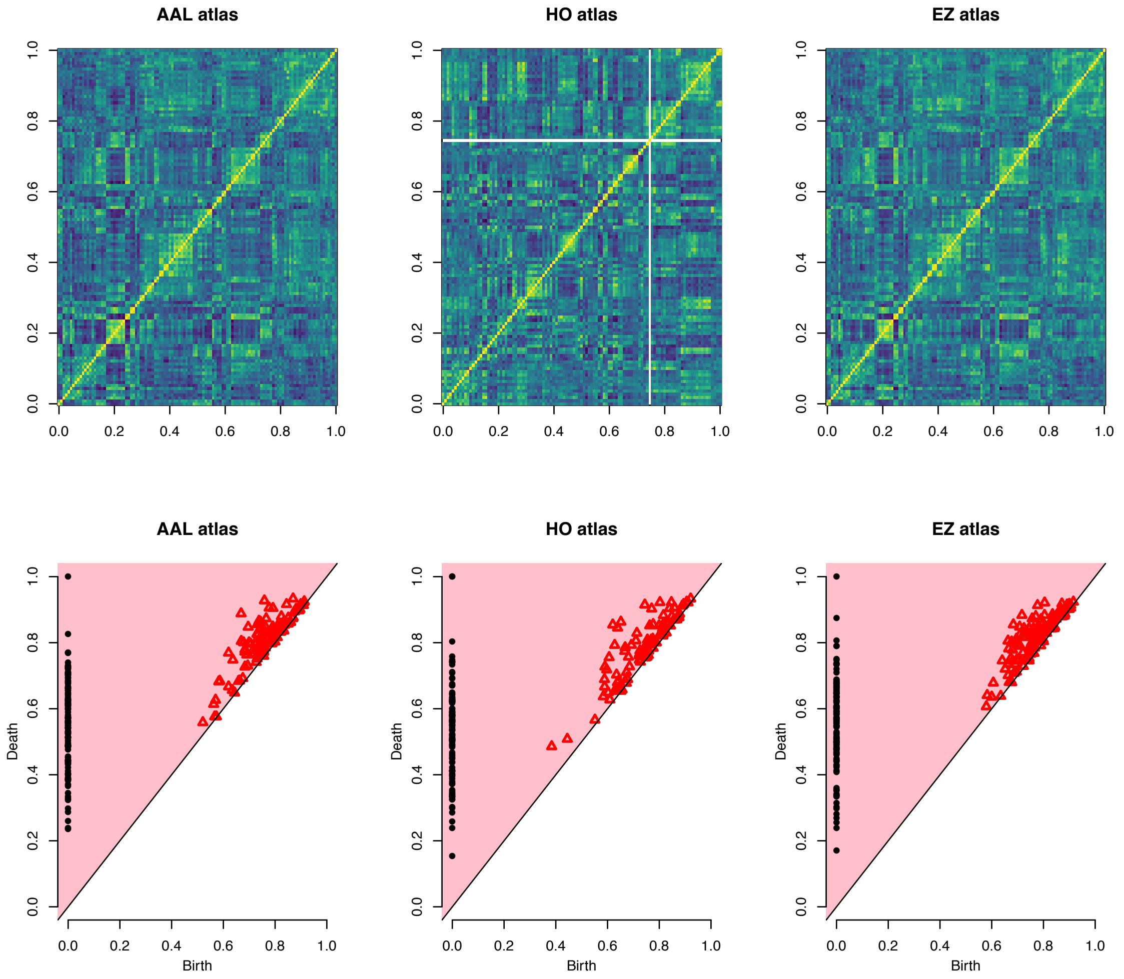

Figure 3.10: Correlation matrices (top) and corresponding Persistence Diagram (bottom) for one subject of the study with closed eyes, with the area of the diagram highlighted in pink representing the confidence band around the diagonal.

Association Measure The second choice to be made to build Persistence Diagram from fMRI data is that of association measure between brain areas, which determines how "close" two regions are. In this preliminary analysis, in which we are more interested in assessing whether there is potential for TDA rather than fully exploit it, we adopt the most basic measure dependency measure, correlation. We summarize the topology of each fMRI scan by computing Persistence Diagrams from the Rips complex defined using as distance

$$d(X_i, X_j) = 1 - (\mathbb{C}_n(X_i, X_j)^2$$

where $\mathbb{C}_n(X_i, X_j)^2$ is the empirical correlation between the fMRI time series corresponding to the $i^{\texttt{th}}$ and $j^{\texttt{th}}$ ROI.

We are aware that a naive correlation-based approach happens to be approximate in describing the functional connectivity, since it neglects the influence of all the other variables and it may lead to nonzero estimates even when the brain regions are independent (Smith et al. 2006) and that several alternative approaches have been investigated to obtain more reliable representations and robust descriptions of the functional networks, such as wavelet based correlation analysis (Achard et al. 2006) and graphical models (Craddock et al. 2013). We plan to investigate this point further in the future.

3.4.2 Phenotipical determinants of Brain Topology

Data analysed in this section are taken from the Autism Brain Imaging Data Exchange (ABIDE) project (`http://preprocessed-connectomes-project.org/abide/index.html`). ABIDE is part of the 1000 Functional Connectome project (`http://fcon_1000.projects.nitrc.org`), a freely accessible collection of resting state fMRI data collected independently by over 33 institution across the world. For our analysis, we consider 794 scans (380 of which correspond to patients affected by autism while the remaining 414 are taken from a healthy control group) taken from 14 institutions. The raw signal was preprocessed using the C-PAC (Configurable Pipeline for the Analysis of Connectome - `https://fcp-indi.github.io`) pipeline for time and motion correction, space registration, intensity normalization and noise removal (see `http://preprocessed-connectomes-project.org/abide/cpac.html` for more details).

An extensive collection of covariates describing the patients' phenotype are additionally provided for each subject (a complete list can be found at `http://fcon_1000.projects.nitrc.org/indi/abide/ABIDE_LEGEND_V1.02.pdf`), however, their use it is strongly compromised by the presence of missing values. Discarding all variables with more than 20% of missing observations, we are left with the only the following 5:

- `age`: age of the subject at time of the scan, raging from
- `diagnosis type`: either 1 for patient with disease or 2 for control subject
- `sex`: gender of the subject, either 1 for male or 2 for female
- `intelligence measures`: IQ scores
- `eye status`: whether the fMRI was recorded with the patient

In order to understand what kind of information Persistence Diagrams retain about brain activity, if any at all, we perform a supervised analysis tackling two different sources of variability, one somehow *constitutional*, while the other more *contingent*. Following the approach described in Section 3.3, we use Topological classification to discriminate between scans recorded with closed eyes and open eyes at first, and then between healthy and patients affected by autism. Average misclassification rates on 10 fold cross validation for the KSVM performed with the Geodesic Laplacian Kernel introduced in Section 3.1 are shown in Figure 3.11.

Our results are encouraging as the Persistence Diagrams seem to be partially able to recover the congenital difference between closed eyes and open eyes scans, while being robust to the presence or not of a disease. Moreover, we see once again that our Kernel performs better than the competition (misclassification rates shown in Figure 3.11 refer only to the classification of eye status, but analogous results holds for the disease detection).

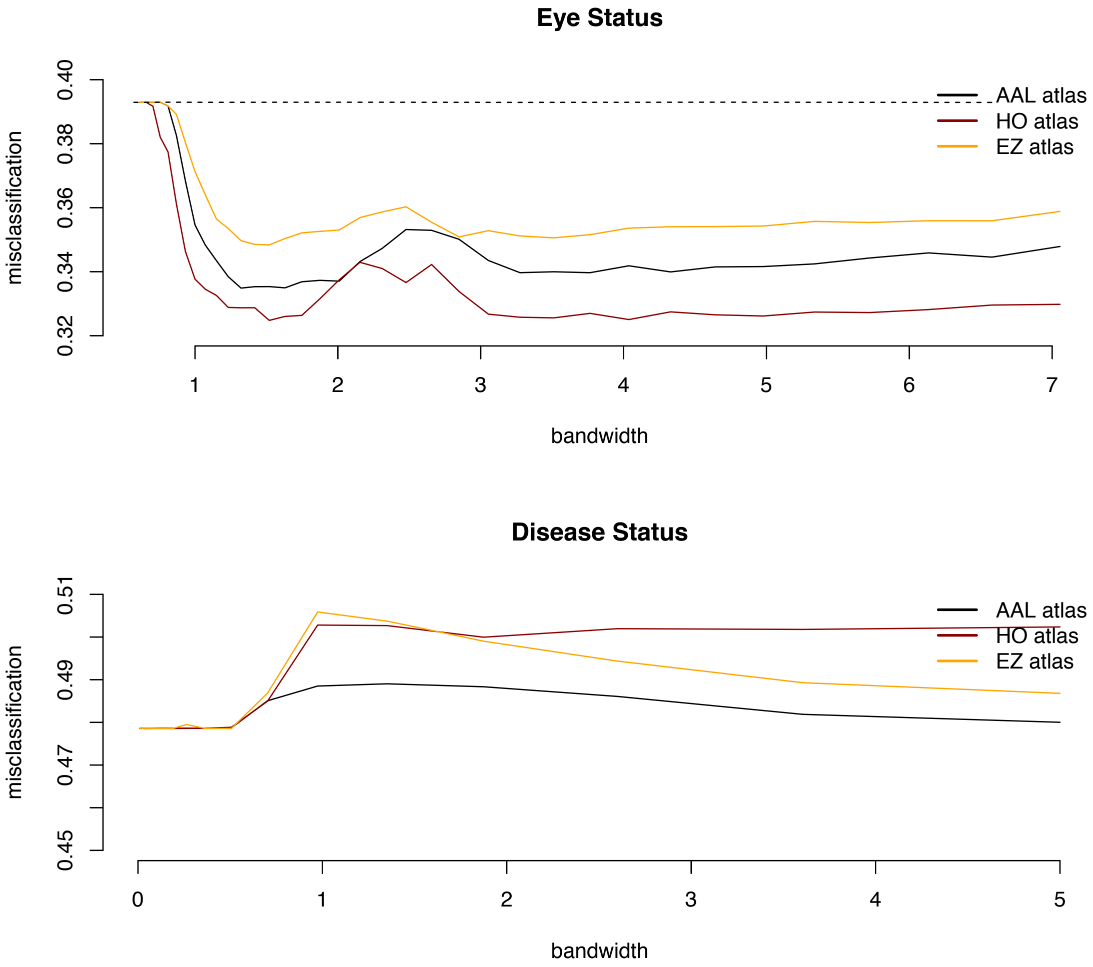

Figure 3.11: Average Misclassification rates for Eye status (top) and Disease status (bottom). Solid lines represent the performance of the Geodesic Laplacian Kernel while the dotten line represents the Persistence Scale Space Kernel.

For our second goal, we now present a strategy to assess how the phenotype affects its composition. We consider as a response variable the number of dimension 1 elements, the loops, of the diagram, which we have just seen retaining some predictive power. The reason we are especially interested in loops is that, while there is a clear interpretation for connected components, in this case, as opposed to the previous examples, it is not equally immediate to understand what higher dimensional topological features may represent. We try to shed some light on this, addressing the problem of assessing phenotypical determinants of topological features. As these are counts data, we model the number of loops in a diagram as a Poisson, and, to keep the notation consistent with the previous chapters, we denote them as Y from here on.

We are aware of the fact that some of the loops that appear in the diagram are not relevant information but they are most likely noisy artifacts, however, in this application trying to clean the diagrams using confidence bands as introduced in Section 2.2.2 turns out to be unsuccessful. Confidence bands may in fact be too conservative and classify as noise even relevant feature (as can be seen in Figure 3.10, where nothing is spared by the band). For this reason we introduce the Degree 1 Total Persistence of the diagram, i.e. the sum of persistence of all features, as Exposure set E. Moreover, the use of the Total Persistence allows to retain part of the information contained in the persistence diagram, while regressing on the number of loops only, would inform us on the determinants of the persistent Betti numbers. While Betti numbers have been shown to be useful in other contexts, Figure 3.12 warns us that in this case they seem to have similar distributions among the two different groups of interest and may not be informative enough.

In order to better approach this issue, we turn, once again, to Quantile regression. The main advantage in doing so, in fact, is that we can assess whether a cleaner diagram (corresponding to a lower level quantile) can actually be better explained, and, if this is the case, we can also see when does the topological noise start to hide the effect of the covariates. On the other hand, if there seems to be signal even at low quantiles, it is a good indication that even less persistent features are actually meaningful. For a fixed quantile level α, our model can be formalized as:

$$\begin{aligned} Y|\lambda &\sim \text{Poisson}(E\lambda) \\ \frac{\lambda}{E} &= \frac{\Gamma^{-1}(q^\alpha + 1, 1-\alpha)}{\Gamma(q^\alpha + 1)} \\ q^\alpha &= \exp \eta^\alpha \end{aligned}$$

where η is the linear predictor, containing fixed and random effect. We compare the following specifications:

- Model 1: $\eta = \beta_0 + \beta_{\texttt{sex}}\texttt{X}_1 + \beta_{age}\texttt{X}_2 + \beta_{\texttt{eye}}\texttt{X}_3 + \beta_{\texttt{diag}}\texttt{X}_4$
- Model 2: $\eta = \beta_0 + \beta_{\texttt{sex}}\texttt{X}_1 + \beta_{age}\texttt{X}_2 + \beta_{\texttt{eye}}\texttt{X}_3 + \beta_{\texttt{diag}}\texttt{X}_4 + \beta_{\texttt{IQ}}\texttt{X}_5$
- Model 3: $\eta = \beta_0 + \beta_{\texttt{sex}}\texttt{X}_1 + \beta_{age}\texttt{X}_2 + \beta_{\texttt{eye}}\texttt{X}_3$
- Model 4: $\eta = \beta_0 + \beta_{\texttt{sex}}\texttt{X}_1 + \beta_{age}\texttt{X}_2 + \beta_{\texttt{eye}}\texttt{X}_3 + \beta_{\texttt{diag}}\texttt{X}_4 + b$
- Model 5: $\eta = \beta_0 + \beta_{\texttt{sex}}\texttt{X}_1 + \beta_{age}\texttt{X}_2 + \beta_{\texttt{eye}}\texttt{X}_3 + \beta_{\texttt{diag}}\texttt{X}_4 + \beta_{\texttt{IQ}}\texttt{X}_5 + b$
- Model 6: $\eta = \beta_0 + \beta_{\texttt{sex}}\texttt{X}_1 + \beta_{age}\texttt{X}_2 + \beta_{\texttt{eye}}\texttt{X}_3 + b$

where b is a Gaussian random effect. In general, the introduction of the random effect seems beneficial to the model, as attested by the lower Deviance Information Criterion (DIC) values (shown in Table 3.6). The main result is that the variable `diagnosis type` does not seem to be significant. Not only the model containing it has worse performance in terms of DIC, but, as shown in Table 3.8 the credibility intervals for $\beta_{\texttt{eye}}$ contain the 0 at every level of the quantile. This indicates some kind of robustness of our topological characterization with respect to the disease condition instead, as already hinted by the poor performance of the classifier shown in Figure 3.11. The variable `eye status` on the other hand, is clearly the one binary variable with the highest impact, which is yet another confirm of the results of the classification.

As the results are robust with respect to the choice of atlas, not only in terms of DIC (Table 3.6), but also in terms of coefficients values, we show only those corresponding to the AAL atlas and Model 4 (Table 3.7 and 3.8). It is worth pointing out that higher level

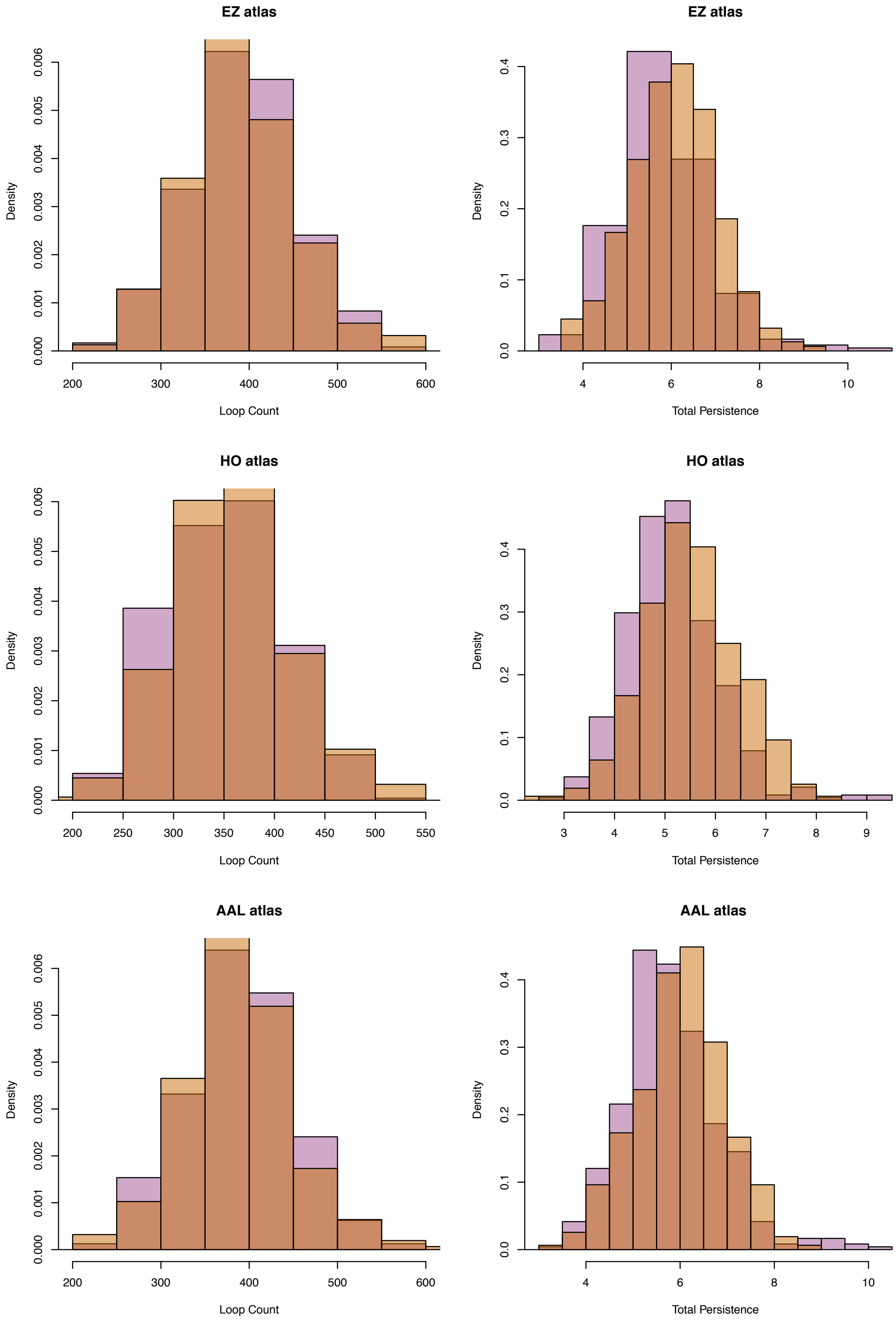

Figure 3.12: Empirical distributions of the Number of Loops per Diagram (left) and Total Persistence (right). Orange corresponds to Open Eye status, while Pink to Closed Eye.

of the quantiles, or diagrams with more features than expected, correspond to estimates for fixed effects coefficients closer to zero. Intuitively this can be seen as a sign that the more components there are, the more noise there is. On the other hand, the coefficients are always larger corresponding to the lower quantiles, meaning that the signal seems to be easier to pick up when the diagram is cleaner and that lower persistent features may in fact be considered noise.

Atlas	Model	$\alpha = 0.05$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 0.95$
AAL	1	11604.45	11602.48	11600.99	11599.43	11597.06
	2	11559.19	11557.07	11555.51	11553.88	11551.44
	3	11605.50	11603.51	11602.02	11600.45	11598.07
	4	7651.30	7651.87	7652.19	7652.47	7652.77
	5	7650.95	7651.47	7651.79	7652.07	7652.37
	6	7651.05	7651.58	7651.91	7652.19	7652.49
HO	1	11709.04	11707.86	11706.86	11705.73	11704.01
	2	11702.87	11701.61	11700.55	11699.38	11697.61
	3	11709.98	11708.80	11707.79	11706.65	11704.92
	4	7580.95	7581.38	7581.63	7581.86	7582.11
	5	7581.21	7581.64	7581.89	7582.11	7582.38
	6	7580.62	7581.05	7581.31	7581.53	7581.79
EZ	1	12238.24	12235.91	12234.17	12232.34	12229.59
	2	12187.43	12184.96	12183.15	12181.26	12178.44
	3	12244.07	12241.73	12239.98	12238.15	12235.40
	4	7668.52	7668.90	7669.13	7669.31	7669.56
	5	7668.40	7668.78	7669.01	7669.19	7669.44
	6	7668.38	7668.76	7669.00	7669.18	7669.43

Table 3.6: DIC values for Model 1-6.

	$\alpha = 0.05$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 0.95$
(Intercept)	3.875063	4.014613	4.104414	4.189111	4.303126
Sex (Female)	0.009358	0.008636	0.008212	0.007839	0.007373
Age	0.008317	0.007779	0.007455	0.007166	0.006798
Diagnosis (Healthy)	-0.003538	-0.003364	-0.003254	-0.003153	-0.003021
Eye Status (Closed)	-0.109502	-0.102413	-0.098150	-0.094330	-0.089480

Table 3.7: Posterior means of fixed effect coefficients of Model 4; for categorical variables the coefficients refer to the variable in the parenthesis, while the reference modalities are, respectively: Male, Autism, Open.

$\alpha = 0.05$	Mean	St. Dev.	0.025 Quant	Median	0.975 Quant	Mode
(Intercept)	3.8750	0.0142	3.8469	3.8750	3.9030	3.8750
Sex (Female)	0.0093	0.0160	-0.0222	0.0093	0.0409	0.0093
Age	0.0083	0.0006	0.0070	0.0083	0.0096	0.0083
Diagnosis (Healthy)	-0.0035	0.0112	-0.0256	-0.0035	0.0186	-0.0035
Eye Status (Closed)	-0.1095	0.0123	-0.1336	-0.1095	-0.0853	-0.1095
$\alpha = 0.25$	Mean	St. Dev.	0.025 Quant	Median	0.975 Quant	Mode
(Intercept)	4.0146	0.0133	3.9883	4.0146	4.0407	4.0146
Sex (Female)	0.0086	0.0150	-0.0208	0.0086	0.0381	0.0086
Age	0.0077	0.0006	0.0065	0.0077	0.0090	0.0077
Diagnosis (Healthy)	-0.0033	0.0105	-0.0240	-0.0033	0.0173	-0.0033
Eye Status (Closed)	-0.1024	0.0115	-0.1249	-0.1024	-0.0798	-0.1024
$\alpha = 0.5$	Mean	St. Dev.	0.025 Quant	Median	0.975 Quant	Mode
(Intercept)	4.1044	0.0127	4.0793	4.1044	4.1294	4.1044
Sex (Female)	0.0082	0.0143	-0.0200	0.0082	0.0364	0.0082
Age	0.0074	0.0005	0.0062	0.0074	0.0086	0.0074
Diagnosis (Healthy)	-0.0032	0.0100	-0.0230	-0.0032	0.0165	-0.0032
Eye Status (Closed)	-0.0981	0.0110	-0.1197	-0.0981	-0.0765	-0.0981
$\alpha = 0.75$	Mean	St. Dev.	0.025 Quant	Median	0.975 Quant	Mode
(Intercept)	4.1891	0.0122	4.1649	4.1891	4.2131	4.1891
Sex (Female)	0.0078	0.0138	-0.0192	0.0078	0.0349	0.0078
Age	0.0071	0.0005	0.0060	0.0071	0.0082	0.0071
Diagnosis (Healthy)	-0.0031	0.0096	-0.0221	-0.0031	0.0158	-0.0031
Eye Status (Closed)	-0.0943	0.0105	-0.1151	-0.0943	-0.0735	-0.0943
$\alpha = 0.95$	Mean	St. Dev.	0.025 Quant	Median	0.975 Quant	Mode
(Intercept)	4.3031	0.0116	4.2802	4.3031	4.3259	4.3031
Sex (Female)	0.0073	0.0130	-0.0183	0.0073	0.0330	0.0073
Age	0.0067	0.0005	0.0057	0.0067	0.0078	0.0067
Diagnosis (Healthy)	-0.0030	0.0091	-0.0210	-0.0030	0.0150	-0.0030
Eye Status (Closed)	-0.0894	0.0100	-0.1091	-0.0894	-0.0698	-0.0894

Table 3.8: Summary of the posterior distribution of fixed effect coefficients of Model 4; for categorical variables the coefficients refer to the variable in the parenthesis, while the reference modalities are, respectively: Male, Autism, Open.

Conclusions

Interpretability has always been a key feature of learning, but it has been even more central in the last years, as a byproduct of the rise of Machine Learning. While (part of) the Statistical community uses this as an argument to distance itself from the overlapping community of Machine Learners, Machine Learning people are quickly catching up. As opposed to most of the literature on interpretable learning, which focuses on interpretable algorithms for learning, we focus on the interpretability of the inputs of such procedures, i.e. on interpretable statistics or summaries of data. We investigated two different approaches of providing accessible characterizations: quantile learning and topological learning.

Quantile Learning The first approach we adopted to provide an intelligible characterization of data was to propose a quantile based parametrization of parametric models. Building on this idea, we recasted Quantile Regression in the more familiar framework of GLM. We explored thoroughly the implication of this link in the Bayesian paradigm, and we showed how the coupling between Quantile Regression and GLM could be exploited in terms of efficiency (thanks to `INLA`) and (posterior) uncertainty quantification. Our second contribution to quantile learning was to provide a model-aware approach to deal with discrete data, thus extending the notion of quantile parametrization beyond continuous models.

Much room is left for improvements. Among the many possible open research question, the first one we plan to investigate is how to borrow strength from the estimation of neighboring quantile in the simultaneous fitting of multiple quantiles. The approach we presented falls in fact into the *Conditional Quantile Models*, where the estimation procedure is carried out separately for each quantile of interest, as opposed to *Joint Quantile Models.* Its advocates, such as Reich, Fuentes, and Dunson (2011) and Tokdar and Kadaney (2012) stress the fact that joint modeling results in ordered quantile curves, hence it is noticeable immune to quantile crossing, however as of now Joint Models have been only explored in the Nonparametric Bayesian framework, which requires stronger assumptions on both covariates and responses, does not allow for linear modeling of the quantiles and it is computationally intensive even when using rough approximations. Our model-based approach could be extended to the case of multiple quantiles by fitting a spline between quantile curves of different levels. thus gaining the advantages of the joint modeling but still keeping a parametric component in the regression model. Additionally, we intend to explore the potential of this new formulation of Quantile Regression in applications, with a special emphasis on Survival Analysis.

Topological Learning In the second part of the thesis we opted for representing data through their connectivity structure, using tools from Topological Data Analysis (TDA), an exciting new field that has seen a tremendous growth in the last couple of years. The theoretical developments have, however, yet to be matched with popularity in applications. In contrast with most of the TDA literature we thus presented a practical framework for this new set of tools.

As far as exploration is concerned, we have introduced the Persistence Flamelet, a new multiscale topological summary, we have characterized it in a probabilistic framework and we have shown how to use it to explore multidimensional time series and the relationship between the bandwidth and the topology of a kernel estimator. In the future we wish to exploit its good probabilistic properties to use it for statistical inference in addition to data description. More specifically, we intend to exploit the CLT for Persistence Flamelets to implement a bootstrap–based testing procedure to assess the significance of topological features and since we characterized the Persistence Flamelets in the context of multivariate time series, we plan to examine their use in testing for change point detection.

Moreover we plan to investigate further the properties of Persistence Flamelets???-related heuristics for bandwidth selection. We have already seen how picking the bandwidth that maximize the persistence seems to be promising, we plan to investigate it even further and to also consider using the Persistence Flamelets to select a bandwidth that stabilizes the topology, in a similar spirit as to Casa and Menardi (2018), possibly by detecting plateau in the Flamelets, or one that reflects some previous knowledge on the topology of the object of interest. Finally, since we can think of the features that appears at many different resolution as the most relevant ones, we also intend to explore persistence in bandwidth ranges as an additional measure of relevance for topological traits.

From a more inferential perspective, we adopted a kernel approach to recast Persistence Diagrams in ready-to-use statistical procedure. We defined a new class of kernels, the Geodesic Topological kernels, which retains more information than other previously defined kernels, and we showed how to exploit them in the context of supervised learning, where their indefiniteness can be easily overcome. Results presented here are encouraging not only for our proposal, which outperforms previously introduced kernels, but for TDA in general. To the best of our knowledge, this is, in fact, the first time that persistence diagrams are used as covariates and highlights the potential of TDA in yet another setting. In the (immediate) future we will release an R-package implementing topological kernels as well as the classification algorithm of Appendix B.

A tale of one thesis In the last part we showed that, in addition to being inspired by the same principles, the set of tools define in this thesis can also benefit from being combined. We illustrated one approach for investigating topological determinants making this, to the best of our knowledge, the first example in which Persistence Diagrams are used as a response variable. Rather than standard regression, we approached the problem by modeling quantiles. This allows for a better control of whether diagram richer in features are also richer in information. While the pipeline provided is well defined, the application we used to introduce it it is still very much work in progress; interestingly, however, we can already elucidate some conclusions. First, and most importantly we can see that there is some predictive power to be exploited in topological characterization per se, and that it is able to capture more "structural" brain activity while being robust to more contingental conditions such as diseases. Secondly, quantile regression does give us a better insight on the noise component in Persistence Diagrams. We now intend to explore more sensible measures of association such as mutual information and distance correlation to properly take into account the temporal component of fMRI signal and add structural connectivity information in order to explore the topological characterization of brain activity to its full potential.

Appendix A

Integrated Nested Laplace Approximation (`INLA`)

The Integrated Nested Laplace Approximation (`INLA`) is a computational method to perform (approximate) Bayesian inference based on Laplace Approximation.

Laplace method is a numerical technique to approximate integrals of the form

$$I = \int f(x)\mathrm{d}x = \int \exp\{\log(f(x))\}\mathrm{d}x.$$

by approximate the target with a Gaussian, matching the mode and the curvature at the mode. The idea is simple and it to use 2nd order Taylor expansion the function in the exponential

$$\begin{aligned}\int f(x)\mathrm{d}x &= \int \exp\{\log(f(x_0))\}\mathrm{d}x \\ &\approx \int \exp\left\{\log(f(x_0)) + \frac{(x-x_0)^2}{2}\frac{\partial^2 \log(f(x))}{\partial x^2}|_{x=x_0}\right\}\end{aligned}$$

where x_0 is the $\arg\max$ of the function f. By setting

$$\sigma = -\frac{1}{\partial^2 \log(f(x))/\partial x^2|_{x=x_0}}$$

we can approximate I with

$$\int f(x)\mathrm{d}x \approx \exp\{\log(f(x))\} \int \exp\left\{-\frac{(x-x_0)^2}{2\sigma^2}\right\}\mathrm{d}x$$

from which we can clearly see a Gaussian Kernel with mean x_0 and variance σ^2.

The reason behind the efficiency of `INLA` is that it is taylored for a specific class of models, Latent Gaussian Models, as opposed to simulation methods, which, in order to be more generic pay a price in efficiency. Formally, the Latent Gaussian models for which `INLA` is designed are hierarchical model of the form:

$$\begin{aligned} y|x,\theta_1 &\sim f(y|x,\theta_1) && \text{Likelihood} \\ x|\theta_2 &\sim \text{GMRF}(\mu,\theta_2) && \text{Latent Field} \\ \theta=(\theta_1,\theta_2) &\sim \pi(\theta) && \text{Hyperpriors} \end{aligned}$$

where the latent field is a Gaussian Markov Random Field (GMRF), whose sparsity contributes to the speed of INLA. Other, less stringent, assumption needed in the INLA approach are:

- the number of hyperparameters $|\theta|$ is small
- the data y are mutually conditionally independent of x and θ, meaning that each observation y_i depends only on one component of the latent field x_i.

Despite the assumption of having a Gaussian Latent field may seem limiting, this is a rather general framework, as there are no restriction on the form of the sampling distribution f nor on the shape of the distribution on the hyperparameter θ, and it includes most common additive models (generalized or not). Additive models are in fact typically defined from a linear predictor of the form

$$\eta_i = \sum_j \beta_j z_{ij} + \sum_k f_k(i)$$

where the β are the coefficient corresponding to the fixed effects, while f_k are model components taking into account spatial or temporal structure, measurement errors and other latent structures in the data. The linear predictor is then related to the likelihood via some link function g, to adapt this framework to any kind of response variable. By assuming the model components f_k to be independet Gaussian processes and by choosing Gaussian priors on the fixed effects β, the latent field $x = (\eta, \beta, f_1, f_2, \dots)$ is a GMRF, with precision given by the sum of the precision matrices of all the components. Notice that depending on the structure of its precision matrix, f_k can be defined as an autoregressive process, smoothing spline, spatial random effect, measurement error correction and so on.

A.1 The Method

The two key quantity INLA is concerned with are the posterior marginals:

$$\pi(x_i|y) = \int \pi(x_i|\theta,y)\pi(\theta|y)d\theta \quad \text{and} \quad \pi(\theta_j|y) = \int \pi(\theta|y)d\theta_{-j}$$

which INLA provides a three step procedure to approximate. More specifically, the INLA algorithm can be specified as follows:

1. Approximate $\tilde{\pi}(\theta|y)$ and get the marginals of the hyperparameters $\tilde{\pi}(\theta_j|y)$ Operationally, the first step in the INLA methodology is to compute an approximation $\tilde{\pi}(\theta|y)$ to $\pi(\theta|y)$, which appears in both Equations.
2. Approximate $\pi(x|\theta, y)$ with another round of Laplace approximation at a given set of points $\theta^{(k)}$

3. Combine the above to get $\tilde{\pi}(x|y)$

$$\tilde{\pi}(x_i|y) = \int \tilde{\pi}(x_i|\theta, y)\tilde{\pi}(\theta|y)d\theta = \sum_k \tilde{\pi}(x_i|\theta^{(k)}, y)\tilde{\pi}(\theta^{(k)}|y)\Delta_k$$

While Step 3 is a rather standard procedure, Step 1 and 2 are the hearth of the `INLA` methodology, hence we provide some additional details.

Step 1 The first step in the `INLA` methodology is to compute an approximation $\tilde{\pi}(\theta|y)$ to $\pi(\theta|y)$, which appears in both Equations. The key relation here is

$$\pi(\theta|y) = \frac{\pi(\theta, x|y)}{\pi(x|y,\theta)} = \frac{\pi(y|x,\theta)\pi(x|\theta)\pi(\theta)}{\pi(y)\pi(x|y,\theta)} \propto \frac{\pi(y|x,\theta)\pi(x|\theta)\pi(\theta)}{\pi(x|y,\theta)}$$

The terms of the numerator are all known, in order to approximate $\pi(\theta|y)$ is enough to substitute the denominator with the Gaussian approximation on the mode (x_*) obtained using Laplace method:

$$\tilde{\pi}(\theta|y) = \left.\frac{\pi(y|x,\theta)\pi(x|\theta)\pi(\theta)}{\tilde{\pi}_G(x|y,\theta)}\right|_{x=x_*} \tag{A.1}$$

In practice, the procedure can be described as:

1. find the mode of $\tilde{\pi}(\theta|y)$ with quasi-Newton methods in order to avoid computation the Hessian at each iteration;
2. compute the Hessian at the modal configuration and invert it to get the matrix Σ;
3. reparametrize θ as
$$\theta(z) = \theta^* + V\Lambda^{1/2}z$$
where $\Sigma = V\Lambda^{1/2}V^{`T}$ is the eigenvalue decomposition of Σ, in order to corrects for scale and location;
4. evaluate $\tilde{\pi}(\theta|y)$ on a grid and then approximate the marginals $\tilde{\pi}(\theta_i|y)$ using an interpolant. Notice that the standardization of the previous step greatly simplifies the choice of grid for the exploration of $\tilde{\pi}(\theta|y)$, as one grid, typically with unit steps can be used for all distributions.

Step 2 The second ingredient needed to compute the marginal posterior of the latent field is $\tilde{\pi}(x_i|\theta, y)$. Rue, Martino, and Chopin (2009) proposes three strategies to do so: a Gaussian approximation, a Laplace approximation and a simplified Laplace approximation.

- Gaussian approximation. This is by far the cheapest in terms of computation, as it takes the marginals of the already computed $\tilde{\pi}_G(x|y,\theta)$ as approximation of $\pi(x_i|y,\theta)$. Since $\tilde{\pi}_G(x|y,\theta)$ is normal then all the marginals will be normal and by exploiting the structure of GMRF it is relatively easy to determine the corresponding means and variances. These results can be computationally very cheap and the approximation often gives reasonable results, but there can be errors in the location or/and errors due to the lack of skewness. This is however a rather coarse approximation, as $\pi(x|y,\theta)$ needs not being Gaussian.

- Laplace approximation. A more accurate way to approximate all marginals is to use Laplace method for each of the marginals separately:

$$\tilde{\pi}_{\mathrm{LA}}(x_i|\theta, y) \propto \left.\frac{\pi(y|x,\theta)\pi(x|\theta)\pi(x)}{\tilde{\pi}_{\mathrm{GG}(x_{-i}|x_i,\theta,y)}}\right|_{x=x_*} \tag{A.2}$$

however, this would require an additional approximation step for the denominator, making the whole procedure computationally intensive. $\tilde{\pi}_{\mathrm{GG}}(x_{-i}|x_i, \theta, y) \propto \pi(y|x, \theta)\pi(x{-}i|\theta, x_i)$ where the first term of the second side is the likelihood which is not normal but can be approximated with the Laplace method. The second term is Gaussian, and thus by using properties of the GMRF we could calculate the corresponding mean and covariance matrix. This implies that $\tilde{\pi}_{\mathrm{GG}}$ must be recomputed for each value of x_i and θ since each precision matrix depends on x_i and θ. This is far too expensive because one needs to repeat all the steps before for the entire latent field (depends on the number of lets say spatial units on the plane).

- Simplified Laplace approximation. This is the default option in `R-INLA` and it is a compromise between the two approaches shown above. Using Taylor expansion of the proper Laplace approximation, it is possible to correct the Gaussian approximation $\tilde{\pi}_G(x|y, \theta)$ for skewness and kurtosis. This strategy is computationally much more convenient than performing a full Laplace approximation but empirically it has shown excellent performances for standard models.

Appendix B

Topological Invariants

In order to better understand what we are trying to recover when estimating topological invariants of some object, it may be useful to recall the basics of topology and what do we mean by topological invariants. According to Klein's definition (Erlangen Program - 1872) topology classifies together objects which can be deformed into one another without cutting or gluing, so basically all the objects that can be deformed into one another without changing the way they are connected. Topology is mostly concerned with connectivity, hence the topological space is the most general space that retains the notion of connectivity.

Definition 1. *Topology* Given a set $\mathbb{X}$ a topology on $\mathbb{X}$ is a collection $\mathcal{O}$ of subsets of $\mathbb{X}$, called open sets, such that:

- $\mathbb{X}$ and $\emptyset \in \mathcal{O}$
- if O_1 and O_2 are open sets, then $O_1 \cap O_2 \in \mathcal{O}$ ($\mathcal{O}$ is closed with respect to the intersection of a finite number of its elements)
- $\mathcal{O}$ is closed with respect to the union of an infinite (possibly uncountable) number of elements.

The topology of a set $\mathbb{X}$ allows us to determine which elements of $\mathbb{X}$ are near, without specifying how distant they are, i.e. without specifying a metric. The pair $(\mathbb{X}, \mathcal{O})$ is known as *Topological space*, a space where each point knows its neighbors. When the topology is clear in the context, we denote the topological space just by $\mathbb{X}$ to keep the notation simple. We have already stated that two topological spaces $\mathbb{X}, \mathbb{Y}$ are equivalent if one can be continuously deformed into the other without changing the way it is connected, i.e. without creating self intersection or holes; the more rigorous definition of this concept is the notion of homeomorphism.

Definition 2. *Homeomorphism* Two topological spaces $\mathbb{X}, \mathbb{Y}$ are *homeomorphic* if there exist a continuous bijection $f : \mathbb{X} \to \mathbb{Y}$ such that its inverse f^{-1} is also continuous. f is called a homeomorphism.

Homeomorphy, denoted by $\approx$, induces the finest level of classification in topology, the topological type. However, it has been proved that is not possible to determine the topological type for spaces whose dimension is larger than 3 (which may be a big limit, when dealing with real data) and so we need to introduce other tools to define equivalence relationship among topological spaces. In practice this tools are maps f, called topological imvariants, that assign the same object to homeomorphic spaces, i.e:

$$\mathbb{X} \approx \mathbb{Y} \Rightarrow f(\mathbb{X}) = f(\mathbb{Y})$$

As we can easily imagine, depending on the choice of f, we can get different topological invariants, ranging from the trivial map, which assigns the same object to every topological space, and thus correspond to the less precise level of classification possible, to the topological type, which is a topological invariant as well, and correspond to the finest level of classification. In between those two extremes, there are infinitely many topological invariants, which may not be as precise as the topological type but are actually computable and are still pretty accurate. We now introduce only a couple of them: homotopy type and homology groups.

In general we say that given two topological spaces $\mathbb{X}, \mathbb{Y}$, two maps $f_0, f_1 : \mathbb{X} \to \mathbb{Y}$ are said to be *homotopic* if there exist a continuous map $F : \mathbb{X} \to \mathbb{Y}$ given by $F(x,t) = f_t(x)$ such that $f_0 = F(x,0)$ and $f_1 = F(x,1)$.

Definition 3. *Homotopy Equivalence* Two topological space $\mathbb{X}, \mathbb{Y}$ are *homotopy equivalent* and have the same homotopy type, denoted by $\mathbb{X} \simeq \mathbb{Y}$ if there exist two continuous maps $f : \mathbb{X} \to \mathbb{Y}$ and $g : \mathbb{Y} \to \mathbb{X}$ such that $g \circ f$ is homotopic to the identity map in $\mathbb{X}$ and $f \circ g$ is homotopic to the identity map in $\mathbb{Y}$.

Classification based on homotopic type is less refined than the one based on the topological type, but homotopy can still be intractable. This is why we introduce an ever coarser topological invariant, the homology groups, which are based on the idea that a topological object can be described in terms of its "holes".

The homology groups of a topological space $\mathbb{X}$ are a set of groups $H_0(\mathbb{X}), H_1(\mathbb{X}), \ldots$, where $H_0(\mathbb{X})$ represents the connected components of $\mathbb{X}$, $H_1(\mathbb{X})$ represents its 1-dimensional cycles and the k-th group $H_k(\mathbb{X})$ represent the k-th dimensional holes of $\mathbb{X}$. Homology is particularly effective in this settings because not only they are easy to compute but also they are discrete by nature (while homeomorphy and homotopy are continuous) and thus it is a better fit for application such as `TDA`, in which the topological information has to be stored in the memory of a computer (which is discrete). %Homology groups are topological invariants since if K and K' are two simplicial complexes with homeomorphic supports $|K| \approx |K'|$ (or even only homotopic), then their homology groups are isomorphic and their Betti numbers are equal. The proof of this is not trivial and relies on the notion of singular homology, which is something more general than

B.1 Simplicial Homology

The easiest way to define homology is to consider the special case of a topological space which consists of simplexes, or special collections of them called simplicial complexes *Simplex* A k-Simplex $\sigma = [v_1, ..., v_k]$ is the convex combination of k affinely independent points $\{v_1, ..., v_k\}$

$$\sigma = \sum_{i=1}^{k} \lambda_i v_i \text{ where } \sum_{i=1}^{k} \lambda_i = 1 \text{ and } \lambda_i \geq 0.$$

The number of affinely independent points, k is dimension of the simplex σ. The convex hull τ of a subset of $\{v_1, ..., v_k\}$ is called a face of σ and we write $\tau \leq \sigma$; if $\tau \neq \sigma$ we say that τ is a *proper face* and we write $\tau < \sigma$. If τ is a proper face of σ, σ is a proper coface of τ.

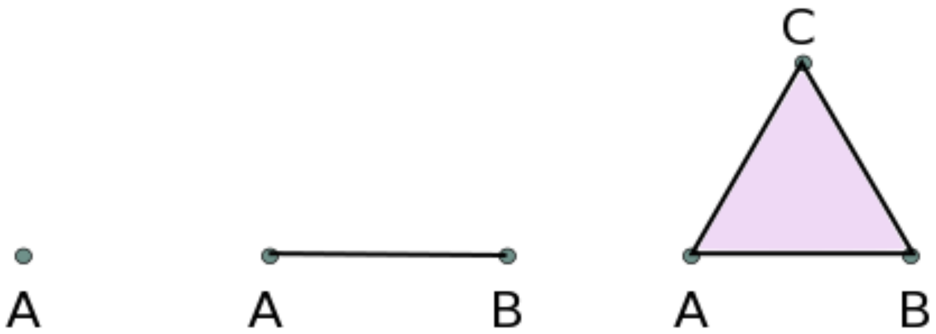

Figure B.1: Example of simplices; a 0-dimensional simplex is a point $\sigma_1 = [A]$, a 1-dimensional simplex is a segment, $\sigma_2 = [A, B]$, a 2-dimensional simplex is a triangle $\sigma_3 = [A, B, C]$ and so on. Notice that the triangle on the right can be also seen as a simplicial complex K if we consider it as the set of all the simplices that form it, i.e. $K = \{[A], [B], [C], [A, B], [A, C], [B, C], [A, B, C]\}$ or alternatively $K = [A] + [B] + [C] + [A, B] + [A, C] + [B, C] + [A, B, C]$

Simplicial complex A simplicial complex is a finite collection of simplices K such that $\sigma \in K$ and $\tau \leq \sigma$ implies $\tau \in K$ and $\sigma, \sigma_0 \in K$ implies $\sigma \cap \sigma_0$ is either empty or a face of both. The dimension of the simplicial complex K is the highest dimension of a simplex belonging to K. Using a simplicial representation it is easier to compute the homology groups of a topological space. It is relatively easy to represent arbitrary objects as a collection of simplicial complexes (this operation is called triangulation [1]), so that simplicial homology is often enough.

Homology describes the topology of a space through a set of finitely generated Abelian groups. Since Abelians groups are commutative groups, we use an additive notation to denote a set of simplices, as shown in Figure B.1.

Chain Given a set of p-simplexes $\{\sigma_1, \ldots, \sigma_m\}$, a p-chain c with coefficients in some ring k is the formal sum:

$$c = \sum_{i=1}^{m} a_i \sigma_i$$

with a_i in the ring k. If we have a simplicial complex K, the space of p-chains generated by the p-simplexes in K is denoted by $C_p(K)$ and it is an Abelian group. A k-dimensional simplicial complex K has a chain group for every dimension $C_1(K), C_2(K), \ldots, C_p(K), \ldots$, even for $p > k$ (in this case the group will be empty, but it is still well defined). In `TDA` the coefficients are in $\mathbb{Z}/2\mathbb{Z} = \{0, 1\}$, hence we can interpret chain groups as sets whose elements are the simplex with non-zero coefficients. Sum between chains is defined component-wise;

[1] In this context a triangulation of a topological space $\mathbb{X}$ is a simplicial complex K such that $\mathbb{X}$ and K are homeomorphic, $\mathbb{X} \approx K$

recall that since in $\mathbb{Z}/2\mathbb{Z}$ we have that $1+1=0$, the sum of two chain is a third chain whose elements are all the simplices which are not in both chains. The most important operator defined on chain groups is the boundary.

Boundary The boundary ∂ of a k-simplex $\sigma = [v_1, ..., v_k]$, is defined as

$$\partial\sigma = \sum_{i=1}^{k} [v_1, ..., \hat{v}_i, ...v_k]$$

where $\hat{v}_i$ means that the vertex v_i is left out of the sum.

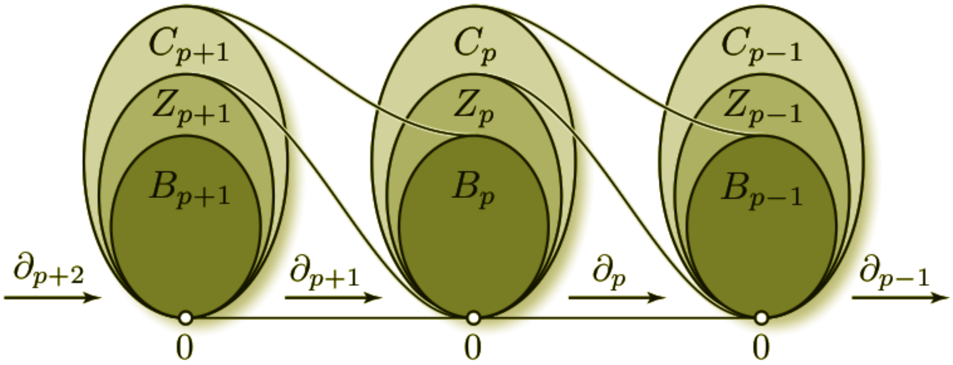

Figure B.2: Chain of complexes (image taken from Edelsbrunner et al. (2010))

The boundary of a k-simplex is the sum of all its $(k-1)$-faces, and so it is a $(k-1)$-chain. In order to underline this dependence on the dimension of the simplex we will denote the boundary for a k-simplex by ∂_k, and so on. More in general we can define the boundary as an homomorphism between chain groups.

Definition 4. *Boundary homomorphism* Let K be a simplicial complex and $\sigma \in K$, $\sigma = [v_1, ..., v_k]$. The boundary homomorphism $\partial_k\sigma : C_k(K) \to C_{k-1}(K)$ is

$$\partial_k\sigma = \sum_{i=1}^{k} [v_1, ..., \hat{v}_i, ...v_k]$$

Notice that in our case, since we define chains with coefficients in $\mathbb{Z}/2\mathbb{Z}$, the chain groups are actually vector spaces, which means that the boundary is a linear operator. As for chains, the boundary homomorphism is also well defined for any dimension (if the chain group is empty, then it will just be the trivial map)

$$0 \to C_k(K) \to C_{k-1}(K) \to ... \to C_0(K) \to 0$$

The boundary ∂_k allows us to define two important subgroups of the chain groups, Cycle groups and Boundary groups.

Definition 5. *Cycle groups* A p-cycle c is a p-chain such that $\partial c = 0$. It immediately follows that, given a simplicial complex K the associated group of p-cycles Z_p is a subgroup of the chain group $C_p(K)$ defined as

$$Z_p(K) = \{c \in C_p(K) | \partial c = 0\}$$

By definition the group of cycles is the kernel of the homomorphism ∂_p, $Z_p = \ker \partial_p$

Definition 6. *Boundary groups* A p-boundary c is the boundary of a $(p+1)$-chain c', $\partial_{p+1}c' = c$. Given a simplicial complex K the associated group of p-cycles B_p is a subgroup of the chain group $C_p(K)$ and by definition the group of p-boundaries is the image of the homomorphism ∂_{p+1}, $B_p = \operatorname{Im} \partial_{p+1}$

Since Cycle and Boundary groups are part of the Chain group, which is Abelian, they are Abelian groups as well. Once again, when we work with coefficients in $\mathbb{Z}/2\mathbb{Z}$ they are vector spaces.

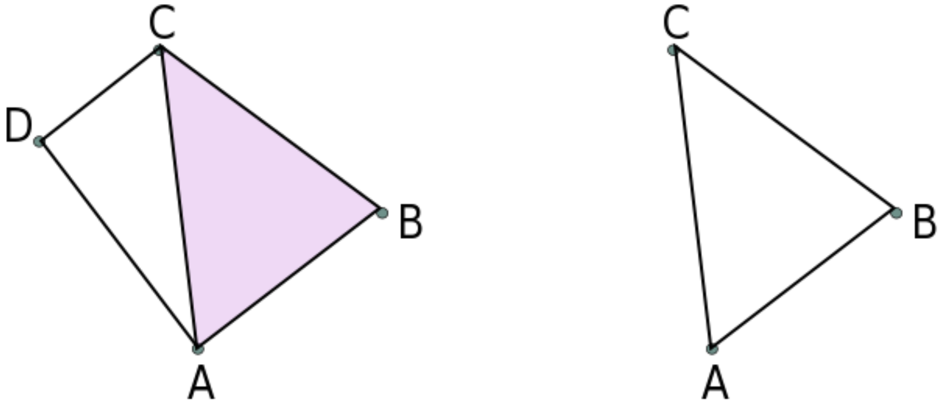

Figure B.3: A 2-Simplicial complex $K = [A] + [B] + [C] + [D] + [A, B] + [A, C] + [B, C] + [B, D] + [C, D] + [A, B, C]$ (on the left) and its 2-boundary, $\partial_2 K$ (on the right).

Example - Boundary Let us consider a simplicial complex K shown in Figure B.3. This is a 2-chain and we want to compute its 2-boundary, ∂_2. By definition we have $\partial_2 K = [A, B] + [A, C] + [B, C]$ (also shown in Figure B.3). $\partial_2 K$ it is still a chain, but of smaller dimension ($\partial_2 K$ is in fact a 1-chain). We can apply the boundary operator again, this time we consider ∂_1. Notice that since the boundary operator is a linear operator we have that for a chain $c = \sum_i \sigma_i$, $\partial c = \sum_i \partial \sigma_i$. $\partial_1 \partial_2 K = \partial[A, B] + \partial[A, C] + \partial[B, C] = [A] + [B] + [A] + [C] + [B] + [C]$. Recall that since we are taking coefficients in $\mathbb{Z}/2\mathbb{Z}$ we have that $[A] + [A] = 0$, and so $\partial_1 \partial_2 K = 0$

The fact that the boundary of a boundary is 0 is not just a coincidence in the previous example, but it is actually the main property of the boundary operator (so much so that it even deserves its own theorem):

Theorem 1. Boundary *For any integer p and for any $(p+1)$-chain c we have that*

$$\partial_p \partial_{p+1} c = 0$$

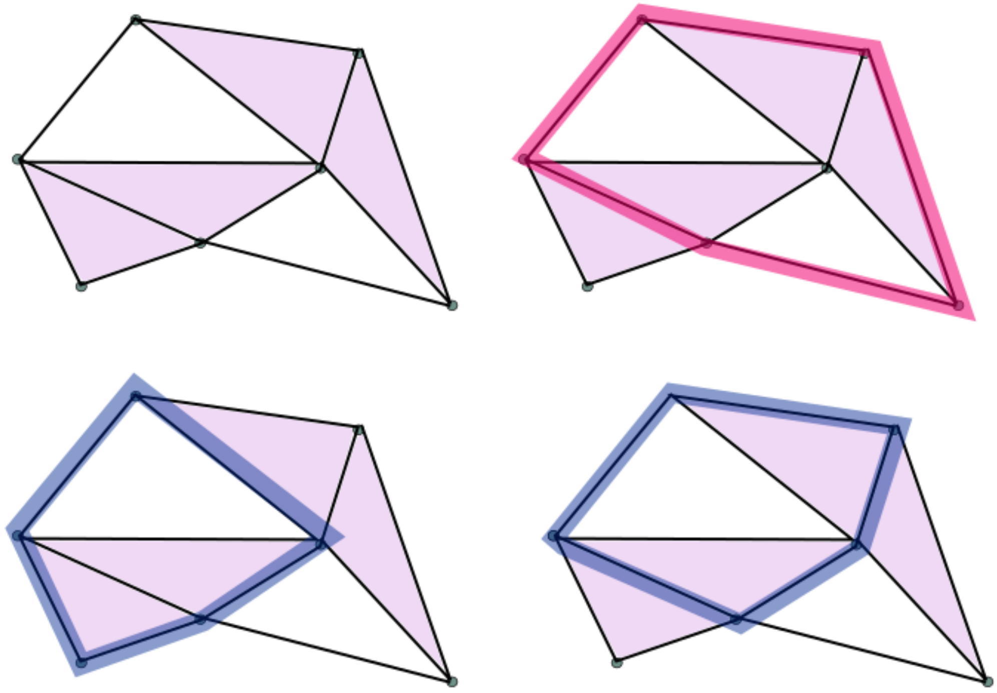

Figure B.4: Homology equivalence. In purple are denoted chains belonging to the same homology group. The chain denoted in pink is instead part of another homology group.

This theorem is crucial because it means that the each boundary it is also a cycle, which allows us to define a quotient group; that quotient group is a homology group.

Definition 7. *Homology groups* The k-th homology group of the simplicial complex K is the quotient group

$$H_k(K) = Z_k/B_k$$

Homology is a map from a topological space to some a sequence of Abelian groups (the homology groups) contained in the groups of chains. Since chain groups are a category, homology can be also defined as a functor, i.e. a map between categories. In the special case where we consider homology groups with coefficients in a field, in our case $\mathbb{Z}_2$, the homology groups are vector spaces.

The rank of homology group is also important and it is called the Betti number:

Definition 8. *Betti number* The k-th Betti number β_k of a simplicial complex K, $\beta_k(K)$ is the rank of the k-homology group of K, i.e.:

$$\beta_k(K) = \operatorname{rank} H_k(K)$$

The Betti number of a simplicial complex K has a straightforward interpretation, the 0-th Betti number represent the number of connected components of K, the 1-st Betti number represents the number of cycles in K, the 2-nd Betti number represents the number of voids and so on, so that the k-th Betti number represents the number of k dimensional "holes" in K. Another way of seeing this is as $\operatorname{rank} H_k = \operatorname{rank} Z_k - \operatorname{rank} B_k$. This is like saying that the k-th Betti number is equal to the difference between the number of simplices that create k cycles and the number of simplices that destroy $k+1$-cycles.

B.1.1 Reduction Algorithm

Homology is usually computed with the so called reduction algorithm. Chain groups with coefficient in $\mathbb{Z}/2\mathbb{Z}$ are vector spaces and so they have a well defined basis. The restriction to coefficients in $\mathbb{Z}/2\mathbb{Z}$ makes things easier to define but it is not necessary; everything we will introduce in the following still holds even if we consider chains with coefficients in an arbitrary ring, since in general the chain groups are finitely generated Abelian groups, which means that they are Abelian groups with a finite basis. Given a basis of C_k, which consists in a set of k-simplices, we can define an integer valued matrix D_k, whose (i,j) element is defined as

$$D_k[i,j] = \begin{cases} 1 & \text{if the } i\text{th } (k-1) \text{ simplex is a face of the } j\text{th } k\text{simplex} \\ 0 & \text{otherwise} \end{cases}$$

that represents the boundary operator $\partial_k : C_k \to C_{k-1}$ with respect to that basis. Each of the columns correspond to one element in the basis of C_k, so that the number of columns of M_k, m_k corresponds to the number of k-simplices in C_k, while the each row correspond to a simplex in the basis of C_{k-1} so that the number of rows, m_{k-1} corresponds to the number of $(k-1)$-simplices in C_k. M_k it is called the standard matrix representation of ∂_k, its null space corresponds to Z_k and its image space to B_{k-1}.

The reduction algorithm consists in reducing D_k to its Smith normal form, $\widetilde{D}_k$,

$$\widetilde{D}_k = \left[\begin{array}{ccc|c} 1 & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & 0 \\ 0 & \dots & 1 & 0 \\ \hline 0 & \dots & 0 & 0 \end{array}\right]$$

using only elementary row and column operations. Each operation on the columns of D_k corresponds to a change in the basis of C_k, while each operation on the row of D_k corresponds to a change in the basis of C_{k-1} so that by reducing the matrix representation of $\underset{\sim}{\partial_k}$ to its diagonal form we find two new bases for C_k and C_{k-1}. The rank of the matrix $\widetilde{D}_k$ The matrix $\widetilde{D}_k$ fully characterizes the homology group H_k, in the sense that:

- The rows with the 1 entries in the diagonal correspond to a basis of B_{k-1}, hence the number of 1 in the diagonal is the rank of the $k-1$ boundary group.
- The columns without 1 entries in the diagonal correspond to a basis of Z_k, hence their number is the rank of the k cycle group.

Recalling that the k-th Betti number is defined as $\beta_k = \operatorname{rank} H_k = \operatorname{rank} Z_k - \operatorname{rank} B_k$, the matrix reduction can be used to compute the Betti numbers.

Appendix C

SVM in RKKS

Given a sample $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$, in its standard formulation in a Reproducing Kernel Hilbert Space $\mathcal{H}$ – i.e. a space generated by a positive definite kernel K – Support Vector Machine (SVM) is defined as the solution to following optimization problem:

$$\begin{cases} \min\limits_{f \in \mathcal{H}, b \in \mathbb{R}} & \frac{1}{2} \|f\|_{\mathcal{H}}^2 = \frac{1}{2} \langle f, f \rangle_{\mathcal{H}}, \\ \text{s.t} & \sum_{i=1}^n \max\left\{0, 1 - y_i \left(f(x_i) + b\right)\right\} \leq \tau, \end{cases} \tag{C.1}$$

or equivalently, in its dual form:

$$\begin{cases} \max\limits_{\boldsymbol{\alpha}} & -\frac{1}{2} \boldsymbol{\alpha}^{\mathsf{t}} \mathbb{G} \, \boldsymbol{\alpha} + \boldsymbol{\alpha}^{\mathsf{t}} \mathbf{1} - \mu \, \boldsymbol{\alpha}^{\mathsf{t}} \boldsymbol{y}, \\ \text{and} & |\alpha_i| \leq \eta, \quad i \in \{1, \ldots, n\}, \end{cases}$$

where $\mathbf{1} \in \mathbb{R}^n$ is a vector of all ones, η is the slack variable and $\mathbb{G}$ the kernel matrix such that $\mathbb{G}_{ij} = y_i \, y_j \, k(x_i, x_j)$.

Extending to the indefinite kernels standard kernel–based classifiers such as Support Vector Machine (SVM) requires some knowledge about Reproducing Kernel Krein Spaces [**?**, **?**]. Every positive kernels are associated to RKHS, similarly each indefinite kernel is associated to a Reproducing Kernel Krein Space (RKKS). A RKKS $\mathcal{K}$ is an inner product space endowed with a Hilbertian topology for which there are two RKHS $\mathcal{K}_+$ and $\mathcal{K}_-$ such that

$$\mathcal{K} = \mathcal{K}_+ \oplus \mathcal{K}_-.$$

RKKS share many properties of RKHS, most noticeably the Riesz and the Representer theorem, which allow to define a solver for the SVM problem.

It has been proven, [**?**], that a minimization problem in a RKHS can be translated into a *stabilization problem* in a RKKS. The SVM optimization problem in a RKKS $\mathcal{K}$ thus can be written as:

$$\begin{cases} \operatorname*{stab}\limits_{f \in \mathcal{K}, b \in \mathbb{R}} & \frac{1}{2} \langle f, f \rangle_{\mathcal{K}} \\ \text{s.t} & \sum_{i=1}^n \max\left\{0, 1 - y_i \left(f(x_i) + b\right)\right\} \leq \tau, \end{cases}$$

which [**?**] proved that can also be written in its dual form

$$\begin{cases} \max\limits_{\widetilde{\boldsymbol{\alpha}}} & -\frac{1}{2} \widetilde{\boldsymbol{\alpha}}^{\mathsf{t}} \widetilde{\mathbb{G}} \, \widetilde{\boldsymbol{\alpha}} + \widetilde{\boldsymbol{\alpha}}^{\mathsf{t}} \mathbf{1} - \mu \, \widetilde{\boldsymbol{\alpha}}^{\mathsf{t}} \boldsymbol{y}, \\ \text{and} & |\widetilde{\alpha}_i| \leq \eta, \quad i \in \{1, \ldots, n\}, \end{cases} \tag{C.2}$$

where $\widetilde{\mathbb{G}} = U\,S\,\Lambda\,U^{\mathtt{t}}$ with U and Λ the eigenvector and eigenvalue matrices of $\mathbb{G} = U\,\Lambda\,U^{\mathtt{t}}$, and $S = \text{sign}(\Lambda)$. Since problem (C.2) is the same as (C.1), it is immediate to see that it can be solved using a standard SVM solver on $\widetilde{\mathbb{G}}$.

References

Achard, Sophie, Raymond Salvador, Brandon Whitcher, John Suckling, and ED Bullmore. 2006. "A Resilient, Low-Frequency, Small-World Human Brain Functional Network with Highly Connected Association Cortical Hubs." *Journal of Neuroscience* 26 (1): 63–72.

Adams, Henry, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. 2017. "Persistence Images: A Stable Vector Representation of Persistent Homology." *J. Mach. Learn. Res.* 18 (1): 218–52.

Alhamzawi, Rahim, and Keming Yu. 2014. "Bayesian Lasso-mixed quantile regression." *Journal of Statistical Computation and Simulation* 84 (4): 868–80.

Alhamzawi, Rahim, Keming Yu, and Dries F Benoit. 2012. "Bayesian Adaptive Lasso Quantile Regression." *Statistical Modelling* 12 (3): 279–97.

Ali, Alnur, J Zico Kolter, and Ryan J Tibshirani. 2016. "The Multiple Quantile Graphical Model." In *Advances in Neural Information Processing Systems*, 3747–55.

Andersen, Anders H, Don M Gash, and Malcolm J Avison. 1999. "Principal Component Analysis of the Dynamic Response Measured by fMRI: A Generalized Linear Systems Framework." *Magnetic Resonance Imaging* 17 (6): 795–815.

Berné, Olivier, and Alexander GGM Tielens. 2012. "Formation of Buckminsterfullerene (C60) in Interstellar Space." *Proceedings of the National Academy of Sciences* 109 (2). National Acad Sciences: 401–6.

Biswal, Bharat B, Maarten Mennes, Xi-Nian Zuo, Suril Gohel, Clare Kelly, Steve M Smith, Christian F Beckmann, et al. 2010. "Toward Discovery Science of Human Brain Function." *Proceedings of the National Academy of Sciences* 107 (10): 4734–9.

Bubenik, Peter. 2015. "Statistical Topological Data Analysis Using Persistence Landscapes." *The Journal of Machine Learning Research* 16 (1): 77–102.

Bullmore, Ed, and Olaf Sporns. 2009. "Complex Brain Networks: Graph Theoretical Analysis of Structural and Functional Systems." *Nature Reviews Neuroscience* 10 (3). Nature Publishing Group: 186–98.

Caillerie, Claire, and Bertrand Michel. 2011. "Model Selection for Simplicial Approximation." *Foundations of Computational Mathematics* 11 (6): 707–31.

Calhoun, Vince Daniel, T Adali, VB McGinty, James J Pekar, TD Watson, and GD Pearlson. 2001. "FMRI Activation in a Visual-Perception Task: Network of Areas Detected Using the General Linear Model and Independent Components Analysis." *NeuroImage* 14 (5): 1080–8.

Carlsson, Gunnar. 2009. "Topology and Data." *Bulletin of the American Mathematical Society* 46 (2): 255–308.

Casa, Alessandro, and Giovanna Menardi. 2018. "Nonparametric Semisupervised Classification for Signal Detection in High Energy Physics." *arXiv Preprint arXiv:1809.02977*.

Chacón, José E, Tarn Duong, and MP Wand. 2011. "Asymptotics for General Multivariate Kernel Density Derivative Estimators." *Statistica Sinica*, 807–40.

Chambers, Ray, Emanuela Dreassi, and Nicola Salvati. 2014. "Disease Mapping via Negative Binomial Regression M-quantiles." *Statistics in Medicine* 33 (27): 4805–24.

Chaudhuri, Probal, and James S Marron. 1999. "SiZer for Exploration of Structures in Curves." *Journal of the American Statistical Association* 94 (447): 807–23.

Chazal, Frédéric, Vin De Silva, Marc Glisse, and Steve Oudot. 2012. "The Structure and Stability of Persistence Modules." *arXiv Preprint arXiv:1207.3674*. Springer.

Chazal, Frédéric, Brittany T Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. 2014a. "Robust Topological Inference: Distance to a Measure and Kernel Distance." *arXiv Preprint arXiv:1412.7197*.

Chazal, Frédéric, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, and Larry Wasserman. 2014b. "Stochastic Convergence of Persistence Landscapes and Silhouettes." In *Proceedings of the Thirtieth Annual Symposium on Computational Geometry*, 474. ACM.

Chazal, Frédéric, Brittany Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, and Larry A Wasserman. 2015. "Subsampling Methods for Persistent Homology." In *ICML*, 2143–51.

Chazal, Frédéric, Marc Glisse, Catherine Labruère, and Bertrand Michel. 2015. "Convergence Rates for Persistence Diagram Estimation in Topological Data Analysis." *The Journal of Machine Learning Research* 16 (1): 3603–35.

Chazal, Frédéric, Leonidas J Guibas, Steve Y Oudot, and Primoz Skraba. 2013. "Persistence-Based Clustering in Riemannian Manifolds." *Journal of the ACM (JACM)* 60 (6): 41.

Cheng, Ming-Yen, Shan Sun, and others. 2006. "Bandwidth Selection for Kernel Quantile Estimation." *Journal of the Chinese Statistical Association* 44 (3): 271–95.

Cohen-Steiner, David, Herbert Edelsbrunner, and Dmitriy Morozov. 2006. "Vines and Vineyards by Updating Persistence in Linear Time." In *Proceedings of the Twenty-Second Annual Symposium on Computational Geometry*, 119–26.

Comaniciu, Dorin, and Peter Meer. 2002. "Mean Shift: A Robust Approach Toward Feature Space Analysis." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (5): 603–19.

Congdon, Peter. 2017. "Quantile Regression for Area Disease Counts: Bayesian Estimation Using Generalized Poisson Regression." *International Journal of Statistics in Medical*

Research 6 (3): 92–103.

Craddock, R Cameron, Saad Jbabdi, Chao-Gan Yan, Joshua T Vogelstein, F Xavier Castellanos, Adriana Di Martino, Clare Kelly, Keith Heberlein, Stan Colcombe, and Michael P Milham. 2013. “Imaging Human Connectomes at the Macroscale.” *Nature Methods* 10 (6): 524–39.

Cristianini, Nello, and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge university press.

Devroye, Luc, and Gary L Wise. 1980. “Detection of Abnormal Behavior via Nonparametric Estimation of the Support.” *SIAM Journal on Applied Mathematics* 38 (3): 480–88.

Edelsbrunner, Herbert, and John Harer. 2010. *Computational Topology: An Introduction.* American Mathematical Soc.

Edelsbrunner, Letscher, and Zomorodian. 2002. “Topological Persistence and Simplification.” *Discrete & Computational Geometry* 28 (4): 511–33.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, and others. 1996. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.” In, 96:226–31.

Falk, Michael. 1984. “Relative Deficiency of Kernel Type Estimators of Quantiles.” *The Annals of Statistics*, 261–68.

Fasiolo, Matteo, Yannig Goude, Raphael Nedellec, and Simon N Wood. 2017. “Fast calibrated additive quantile regression.” *arXiv Preprint arXiv:1707.03307.*

Fasy, Brittany T, Jisu Kim, Fabrizio Lecci, Clement Maria, and V Rouvreau. 2014. “TDA: Statistical Tools for Topological Data Analysis.” *Availabl E at Https://Cran. R-Project. Org/Web/Packages/TDA/Index. Html.*

Fasy, Brittany Terese, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, Aarti Singh, and others. 2014. “Confidence Sets for Persistence Diagrams.” *The Annals of Statistics* 42 (6): 2301–39.

Feragen, Aasa, Francois Lauze, and Soren Hauberg. 2015. “Geodesic Exponential Kernels: When Curvature and Linearity Conflict.” In *Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition*, 3032–42.

Galton, Francis. 1883. *Inquiries into human faculty and its development.* JM Dent; Company.

Genovese, Christopher R, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. 2016. “Non-Parametric Inference for Density Modes.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78 (1): 99–126.

Genovese, Christopher, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. 2017. “Finding Singular Features.” *Journal of Computational and Graphical Statistics.*

Taylor & Francis, 1–12.

Geraci, Marco. 2017. "Nonlinear quantile mixed models." *arXiv Preprint arXiv:1712.09981*.

Geraci, Marco, and Matteo Bottai. 2007. "Quantile regression for longitudinal data using the asymmetric Laplace distribution." *Biostatistics* 8 (1): 140–54.

———. 2014. "Linear quantile mixed models." *Statistics and Computing* 24 (3): 461–79.

Gilchrist, Warren. 2008. "Regression revisited." *International Statistical Review* 76 (3): 401–18.

Godtliebsen, Fred, James Stephen Marron, and Probal Chaudhuri. 2004. "Statistical Significance of Features in Digital Images." *Image and Vision Computing* 22 (13): 1093–1104.

Gretton, Arthur, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. "A Kernel Two-Sample Test." *Journal of Machine Learning Research* 13 (Mar): 723–73.

Groeneveld, Richard A, and Glen Meeden. 1984. "Measuring Skewness and Kurtosis." *The Statistician*, 391–99.

Hatcher, Allen. 2002. *Algebraic Topology*. Cambridge university press.

Härdle, Wolfgang. 1990. *Applied Nonparametric Regression*. Cambridge university press.

Härdle, Wolfgang Karl, Marlene Müller, Stefan Sperlich, and Axel Werwatz. 2012. *Nonparametric and Semiparametric Models*. Springer Series in Statistics. Springer.

Ilienko, Andrii. 2013. "Continuous counterparts of Poisson and binomial distributions and their properties." *arXiv Preprint arXiv:1303.5990*.

Jones, MC. 2008. "On a Class of Distributions with Simple Exponential Tails." *Statistica Sinica*, 1101–10.

Koenker, Roger W, and Vasco d'Orey. 1987. "Computing Regression Quantiles." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 36 (3): 383–93.

Koenker, Roger, and Gilbert Bassett. 1978. "Regression Quantiles." *Econometrica* 46 (1): 33–50.

Koenker, Roger, and Pin Ng. 2005. "A Frisch-Newton Algorithm for Sparse Quantile Regression." *Acta Mathematicae Applicatae Sinica* 21 (2): 225–36.

Kozumi, Hideo, and Genya Kobayashi. 2011. "Gibbs sampling methods for bayesian quantile regression." *Journal of Statistical Computation and Simulation* 81 (11): 1565–78.

Kraus, Daniel, and Claudia Czado. 2017. "D-vine copula based quantile regression." *Computational Statistics and Data Analysis* 110: 1–18.

Kusano, Genki, Yasuaki Hiraoka, and Kenji Fukumizu. 2016. "Persistence Weighted Gaussian Kernel for Topological Data Analysis." In *International Conference on Machine*

Learning, 2004–13.

Ledoux, Michel, and Michel Talagrand. 2013. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer.

Lee, Megan H, Christopher D Smyser, and Joshua S Shimony. 2013. "Resting-State fMRI: A Review of Methods and Clinical Applications." *American Journal of Neuroradiology* 34 (10): 1866–72.

Li, Qi, and Jeffrey Scott Racine. 2007. *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Li, Qing, Ruibin Xiy, and Nan Lin. 2010. "Bayesian regularized quantile regression." *Bayesian Analysis* 5 (3): 533–56.

Loosli, Gaëlle, Stéphane Canu, and Cheng Soon Ong. 2016. "Learning Svm in Krein Spaces." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38 (6): 1204–16.

Lum, Kristian, and Alan E. Gelfand. 2012. "Spatial quantile multiple regression using the asymmetric Laplace process." *Bayesian Analysis* 7 (2): 235–58.

Luss, Ronny, and Alexandre d'Aspremont. 2008. "Support Vector Machine Classification with Indefinite Kernels." In *Advances in Neural Information Processing Systems*, 953–60.

Ma, Yanyuan, Marc G. Genton, and Emanuel Parzen. 2011. "Asymptotic properties of sample quantiles of discrete distributions." *Annals of the Institute of Statistical Mathematics* 63 (2): 227–43.

Machado, José A.F., and J. M.C. Santos Silva. 2005. "Quantiles for counts." *Journal of the American Statistical Association* 100 (472): 1226–37.

Marino, Maria Francesca, and Alessio Farcomeni. 2015. "Linear quantile regression models for longitudinal experiments: An overview" 73 (2): 229–47.

Meinshausen, Nicolai. 2006. "Quantile Regression Forests." *Journal of Machine Learning Research* 7: 983–99.

Mileyko, Yuriy, Sayan Mukherjee, and John Harer. 2011. "Probability Measures on the Space of Persistence Diagrams." *Inverse Problems* 27 (12): 124007.

Moon, Chul, Noah Giansiracusa, and Nicole A. Lazar. 2018. "Persistence Terrace for Topological Inference of Point Cloud Data." *Journal of Computational and Graphical Statistics* 27 (3): 576–86.

Morozov, Dmitriy. 2008. *Homological Illusions of Persistence and Stability*. Duke University.

———. 2012. "Dionysus." *Software Available at Http://Www.mrzv. org/Software/Dionysus*.

Munch, Elizabeth. 2013. "Applications of Persistent Homology to Time Varying Systems." PhD thesis, Duke University.

Munch, Elizabeth, Katharine Turner, Paul Bendich, Sayan Mukherjee, Jonathan Mattingly, John Harer, and others. 2015. "Probabilistic Fréchet Means for Time Varying Persistence

Diagrams." *Electronic Journal of Statistics* 9 (1): 1173–1204.

Perea, Jose A, and John Harer. 2015. "Sliding Windows and Persistence: An Application of Topological Methods to Signal Analysis." *Foundations of Computational Mathematics* 15 (3). Springer: 799–838.

Reich, Brian J., Montserrat Fuentes, and David B. Dunson. 2011. *Journal of the American Statistical Association* 106 (493): 6–20.

Reininghaus, Jan, Stefan Huber, Ulrich Bauer, and Roland Kwitt. 2015. "A Stable Multi-Scale Kernel for Topological Machine Learning." In *Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition*, 4741–8.

Riebler, Andrea, Sigrunn H Sørbye, Daniel Simpson, and Håvard Rue. 2016. "An intuitive Bayesian spatial model for disease mapping that accounts for scaling." *Statistical Methods in Medical Research* 25 (4): 1145–1116.

Rue, Håvard, and Leonhard Held. 2005. *Gaussian Markov Random Fields Theory and Applications*. CRC press.

Rue, Håvard, Sara Martino, and Nicolas Chopin. 2009. "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 71 (2): 319–92.

Rue, Håvard, Andrea Riebler, Sigrunn H Sørbye, Janine B Illian, Daniel P Simpson, and Finn K Lindgren. 2017. "Bayesian Computing with INLA : A Review." *Annual Review of Statistics and Its Application* 4: 395–421.

Santos, Damiana A, and Marcos Duarte. 2016. "A Public Data Set of Human Balance Evaluations." *PeerJ* 4: e2648.

Scholkopf, Bernhard, and Alexander J Smola. 2001. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press.

Scott, David W. 2015. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons.

Simpson, Daniel, Håvard Rue, Andrea Riebler, Thiago G Martins, and Sigrunn Holbek Sørbye. 2017. "Penalising model component complexity: A principled, practical approach to constructing priors." *Statistical Science* 32 (1): 1–28.

Smith, Stephen M, Mark Jenkinson, Heidi Johansen-Berg, Daniel Rueckert, Thomas E Nichols, Clare E Mackay, Kate E Watkins, et al. 2006. "Tract-Based Spatial Statistics: Voxelwise Analysis of Multi-Subject Diffusion Data." *Neuroimage* 31 (4): 1487–1505.

Sriram, Karthik. 2015. "A sandwich likelihood correction for Bayesian quantile regression based on the misspecified asymmetric Laplace density." *Statistics and Probability Letters* 107: 18–26. `http://dx.doi.org/10.1016/j.spl.2015.07.035`.

Sriram, Karthik, R. V. Ramamoorthi, and Pulak Ghosh. 2013. "Posterior consistency of bayesian quantile regression based on the misspecied asymmetric laplace density." *Bayesian Analysis* 8 (2): 479–504. doi:10.1214/13-BA817.

Sørbye, Sigrunn Holbek, and Håvard Rue. 2014. "Scaling intrinsic Gaussian Markov random field priors in spatial modelling." *Spatial Statistics* 8: 39–51.

Takeuchi, Ichiro, Quoc V. Le, Tim Sears, and Alexander J. Smola. 2005. "Nonparametric Quantile Regression." *Journal of Machine Learning Research* 7.

Tokdar, Surya T., and Joseph B. Kadaney. 2012. "Simultaneous linear quantile regression: A semiparametric bayesian approach." *Bayesian Analysis* 7 (1): 51–72.

Turner, Katharine, Yuriy Mileyko, Sayan Mukherjee, and John Harer. 2014. "Fréchet Means for Distributions of Persistence Diagrams." *Discrete & Computational Geometry* 52 (1): 44–70.

Van Dijk, Koene RA, Trey Hedden, Archana Venkataraman, Karleyton C Evans, Sara W Lazar, and Randy L Buckner. 2009. "Intrinsic Functional Connectivity as a Tool for Human Connectomics: Theory, Properties, and Optimization." *Journal of Neurophysiology* 103 (1): 297–321.

Wang, Huixia Judy, Ian W. McKeague, and Min Qian. 2017. "Testing for marginal linear effects in quantile regression." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80 (2): 433–52.

Wang, Yuan, Hernando Ombao, and Moo K Chung. 2018. "Topological Data Analysis of Single-Trial Electroencephalographic Signals." *The Annals of Applied Statistics* 12 (3). NIH Public Access: 1506.

Yan, Yifei, and Athanasios Kottas. 2017. "A New Family of Error Distributions for Bayesian Quantile Regression." *arXiv Preprint arXiv:1701.05666.*

Yang, Yunwen, Huixia Judy Wang, and Xuming He. 2016. "Posterior Inference in Bayesian Quantile Regression with Asymmetric Laplace Likelihood." *International Statistical Review* 84 (3): 327–44.

Yu, Keming, and M C Jones. 1998. "Local Linear Quantile Regression Local Linear Quantile Regression." *Journal of the American Statistical Association* 93 (441): 228–37.

Yu, Keming, and Rana Moyeed. 2001. "Bayesian quantile regression." *Statistics & Probability Letters* 54: 437–47.

Yue, Yu Ryan, and Håvard Rue. 2011. "Bayesian inference for additive mixed quantile regression models." *Computational Statistics and Data Analysis* 55 (1): 84–96.